

# Alternative Tests for Monotonicity in Expected Asset Returns\*

Joseph P. Romano<sup>†</sup>

Departments of Statistics and Economics  
Stanford University  
Sequoia Hall  
Stanford, CA 94305, USA  
romano@stanford.edu

Michael Wolf<sup>‡</sup>

Department of Economics  
University of Zurich  
Bluemlisalpstrasse 10  
CH-8006 Zurich, Switzerland  
michael.wolf@econ.uzh.ch

May 2011

## Abstract

Many postulated relations in finance imply that expected asset returns should monotonically increase in a certain characteristic. To examine the validity of such a claim, one typically considers a finite number of return categories, ordered according to the underlying characteristic. A standard approach is to simply test for a difference in expected returns between the highest and the lowest return category. However, such an approach can be misleading, since the relation of expected returns could be flat, or even decreasing, in the range of intermediate categories. A new test, taking the entire range of categories into account, has been proposed by [Patton and Timmermann \(2010\)](#). Unfortunately, the test is based on an additional assumption that can be violated in many applications of practical interest. As a consequence, it can be quite likely for the test to ‘establish’ strict monotonicity of expected asset returns when such a relation actually does not exist. We offer some alternative tests which do not share this problem. The behavior of the various tests is illustrated via Monte Carlo studies. We also present empirical applications to real data.

KEYWORDS: Bootstrap; CAPM; Monotonicity tests; Systematic relation.

JEL CLASSIFICATION NUMBERS: C12, C58, G12, G14.

---

\*We thank Olivier Ledoit, Markus Leippold, and various seminar participants for helpful discussions.

<sup>†</sup>Research has been supported by NSF Grant DMS-0707085.

<sup>‡</sup>Corresponding author. Phone: +41-44 634 5096; Fax: +41-44 634 4907. Research has been supported by the Swiss National Science Foundation (NCCR FINRISK, Module A3).

# 1 Introduction

Many postulated relations in finance imply that expected asset returns should monotonically increase (or decrease) in a certain characteristic. When the assets are equities, examples of such characteristics are CAPM beta, book-to-market, size, momentum, and reversal. When the assets are bonds, examples of such characteristics are maturity and rating quality. The search for new such characteristics is a never-ending quest, partly in hopes of creating novel trading strategies to ‘beat the market’.

It is, therefore, of interest to test whether a particular characteristic indeed generates expected asset returns that are monotonically increasing (or decreasing). Say there are total of  $N + 1$  categories, ordered according to the underlying characteristic. The postulated relation says that as one moves up from one category to the next, then the expected return should strictly increase (or decrease). From here on, we will only focus on the case of expected asset returns being supposedly monotonically increasing. The opposite case of expected asset returns being supposedly monotonically decreasing can be handled analogously by simply multiplying all returns by negative one or, alternatively, by reversing the order of the various return categories considered.

For a long time, the standard in the field has been to simply test for a difference in expected returns between the highest and the lowest return category. Such a test is easily carried out, since the parameter of interest is univariate, being the difference between two expected values. Therefore, a conventional  $t$ -test can be applied, though one has to account for potential serial correlation of returns in computing the standard error that appears in the denominator of the test statistic.

A new test has been proposed recently by [Patton and Timmermann \(2010\)](#), abbreviated by P&T henceforth. As they point out, simply testing for a difference in expected returns between the highest and the lowest category can be misleading. It could happen that the relation of expected returns is flat, or even decreasing, for intermediate return categories while a positive difference ‘highest minus lowest’ still exists. In this case, providing sufficient data are collected, the simple  $t$ -test is very likely to falsely decide in favor of a monotonic relation. Take the example of five return categories, ordered from lowest to highest, with respective expected returns of 1.0, 1.5, 1.1, 1.4 and 1.6. In this example, the overall relation is not monotonic, even though the difference ‘highest minus lowest’ is positive.

Therefore, a more comprehensive approach is needed to establish a monotonic relation over the entire range of return categories. When moving from one category up to the next, the difference in expected returns must be established as significantly positive every time. In other words, all  $N$  expected return differentials ‘higher minus lower’ must be established as significantly greater than zero. Such a test is more complex, since the underlying parameter

is now an  $N$ -dimensional vector rather than a univariate number.

A natural test statistic for the more comprehensive approach is obtained as follows. Compute individual  $t$ -test statistics for each of the  $N$  expected return differentials, where in each case the alternative corresponds to the expected return differential being positive. Then take the minimum of the individual test statistics as the overall test statistic. If the resulting min- $t$  statistic is ‘sufficiently’ large, one decides in favor of a strictly monotonic relation. The statistical question is how to obtain a proper critical value for this test. P&T use a bootstrap method, resampling from a certain least favorable null distribution. Unfortunately, a closer inspection reveals that the proposed test is only valid, in terms of controlling the probability of a type 1 error, when making an additional assumption. The additional assumption is that if the relation is not strictly monotonically increasing, it must be weakly monotonically decreasing. That is, if the expected return differentials are not all strictly positive, then they must be all weakly negative (meaning less than or equal to zero).

While this additional assumption may hold for certain applications, it certainly cannot be invoked whenever testing for monotonicity of expected asset returns. When it is possible a priori that some expected return differentials can be strictly positive while others can be weakly negative, then the additional assumption is unreasonable. In this more general case, the P&T test is not successful at controlling the rejection probability under the null. It can decide in favor of the alternative of a monotonic relation with quite a large probability when in fact such a relation does not exist. In this paper, we first discuss this problem of the P&T test and then offer some alternative tests that do not share the problem and are, therefore, safer to use in the general scenario of a relation that might not be systematic.

Having said this, all the new tests we present have the choice of alternative hypothesis in common with the P&T test. The alternative postulates that a monotonic relation (in the sense of being strictly increasing) exists. In other proposals, this postulate becomes the null hypothesis instead; for example, see [Wolak \(1987, 1989\)](#) and also [Fama \(1984\)](#). Such an approach is unnatural, though, if the goal is to establish the existence of a monotonic relation. By not rejecting the null, one can never claim statistical evidence in favor of the associated hypothesis. Hence, we will not include such tests in our paper. For a theoretical discussion of such tests and some examination of finite-sample performance, see P&T.

The remainder of the paper is organized as follows. [Section 2](#) describes the formal setup and the testing problem of interest. [Section 3](#) presents various approaches for designing tests for monotonicity. [Section 4](#) details how the various tests are implemented in practice. [Section 5](#) examines finite-sample performance via Monte Carlo studies. [Section 6](#) contains empirical applications to real-life data. Finally, [Section 7](#) concludes. Mathematical results as well as all figures and tables are relegated to the Appendix.

## 2 Formal set-up and testing problem

Our notation generally follows the notation of P&T. One observes a strictly stationary time series of return vectors  $\mathbf{r}_t \equiv (r_{t,0}, r_{t,1}, \dots, r_{t,N})'$  of dimension  $N + 1$ . The observation period runs from  $t = 1$  to  $t = T$ , so the sample size is given by  $T$ . Denote the expected return vector by  $\boldsymbol{\mu} \equiv (\mu_0, \mu_1, \dots, \mu_N)'$  and define the associated expected return differentials as

$$\Delta_i \equiv \mu_i - \mu_{i-1} \quad \text{for } i = 1, \dots, N. \quad (1)$$

To avoid any possible confusion, note that the characteristic according to which the  $N + 1$  return categories are ordered is always assumed to be predetermined and not data-dependent. We also introduce the following notation for the observed return differentials:

$$\mathbf{d}_t \equiv (d_{t,1}, \dots, d_{t,N})' \equiv (r_{t,1} - r_{t,0}, \dots, r_{t,N+1} - r_{t,N})'. \quad (2)$$

Therefore, letting  $\boldsymbol{\Delta} \equiv (\Delta_1, \dots, \Delta_N)'$ , one can also write  $\boldsymbol{\Delta} = \mathbb{E}(\mathbf{d}_t)$ .

The approach proposed by P&T specifies a completely flat or weakly decreasing pattern under the null and a strictly increasing pattern under the alternative:

$$H_0 : \Delta_i \leq 0 \text{ for all } i \quad \text{vs.} \quad H_1 : \Delta_i > 0 \text{ for all } i. \quad (3)$$

Alternatively, these hypotheses can be expressed as<sup>1</sup>

$$H_0 : \boldsymbol{\Delta} \leq \mathbf{0} \quad \text{vs.} \quad H_1 : \boldsymbol{\Delta} > \mathbf{0}. \quad (4)$$

And yet another equivalent expression is

$$H_0 : \Delta_i \leq 0 \text{ for all } i \quad \text{vs.} \quad H_1 : \min_i \Delta_i > 0. \quad (5)$$

The problem with this approach is that the alternative is not the negation of the null, at least if the parameter space for  $\boldsymbol{\Delta}$  is  $\mathbb{R}^N$ , the  $N$ -dimensional Euclidian space. To see this, one can partition  $\mathbb{R}^N$  as follows:

$$\mathbb{R}^N = R_1 \cup R_2 \cup R_3, \quad (6)$$

with

$$R_1 \equiv \{\mathbf{x} \in \mathbb{R}^N : x_i \leq 0 \text{ for all } i\},$$

$$R_2 \equiv \{\mathbf{x} \in \mathbb{R}^N : x_i \leq 0 \text{ for some } i \text{ and } x_j > 0 \text{ for some other } j\},$$

$$\text{and } R_3 \equiv \{\mathbf{x} \in \mathbb{R}^N : x_i > 0 \text{ for all } i\}.$$

For example, when  $N = 2$ , then  $R_1$  corresponds to the third quadrant (including the axes),  $R_3$  corresponds to the first quadrant (excluding the axes), and  $R_2$  corresponds to the rest of

---

<sup>1</sup>Equalities and inequalities for vectors are interpreted as applying component by component.

the real plane (that is, the second plus the fourth quadrant with some axes included and others excluded).

The hypotheses of (3)–(5) can then also be expressed as

$$H_0 : \Delta \in R_1 \quad \text{vs.} \quad H_1 : \Delta \in R_3 . \quad (7)$$

From the particular expression (7), it is now easily appreciated that the alternative is the negation of the null only under the *additional* assumption

$$\Delta \in R_1 \cup R_3 . \quad (8)$$

In other words, it is ruled out a priori that  $\Delta \in R_2$ , namely that some expected return differentials are less than or equal to zero while others are positive at the same time. In general, such an assumption is rather restrictive. When  $N = 2$ , then  $R_2$  already makes up half of the entire space  $\mathbb{R}^N$ . But as  $N$  increases, the fraction of  $R_2$  as part of  $\mathbb{R}^N$  gets even larger. For general  $N$ , this fraction equals  $1 - 0.5^{N-1}$ , which quickly approaches one as  $N$  increases. To be fair, it should be pointed out that most of  $R_2$  is not really of concern, only the part of  $R_2$  that is ‘close’ to  $R_3$ ; see Sections 4 and 5.

To sum up, the P&T test reveals whether a null hypothesis of a weakly monotonic decreasing relation, as opposed to a null hypothesis of no systematic relation, can be rejected in favor of a strictly monotonic increasing relation.

If the goal is to establish (strict) monotonicity of expected asset returns and one allows for  $\Delta \in \mathbb{R}^N$  a priori, then the hypotheses must be formulated as follows instead:

$$H_0 : \Delta \in R_1 \cup R_2 \quad \text{vs.} \quad H_1 : \Delta \in R_3 \quad (9)$$

and can be also be expressed as

$$H_0 : \min_i \Delta_i \leq 0 \quad \text{vs.} \quad H_1 : \min_i \Delta_i > 0 . \quad (10)$$

In this more general formulation, the null hypothesis indeed corresponds to no systematic relation.

The distinction between (7) and (9) is of crucial importance for the proper implementation of the test. This will be explained in detail in the next section.

**Remark 2.1** Postulated relations that imply strictly increasing expected returns can be based either on economic theory or on empirical findings (that is, so-called market anomalies). In the context of equity returns, an example of the former is the CAPM while examples of the latter are book-to-market, size, momentum, and reversal.<sup>2</sup>

---

<sup>2</sup>There may exist certain models that also imply monotonicity of expected returns for some of the latter examples. Arguably, such models have often been derived *after* observing market anomalies in data. As a result, they do not enjoy the standing of the CAPM in the finance community.

The CAPM postulates that the expected (excess) return of a stock is a linear function of its CAPM beta, with the slope of the linear function being positive. Assume now that stocks are grouped into deciles, say, based on their CAPM beta. Then, if the linearity of the postulated relation is taken for granted, the expected returns of the ten categories will be either strictly increasing (namely if the slope of the linear relation is indeed positive) or weakly decreasing (namely if the slope of the linear relation is non-positive instead). In this case, the correct formulation of the testing problem is given by (7). On the other hand, if the linearity of the postulated relation is not taken for granted, the correct formulation of the testing problem is given by (9) instead.

The same reasoning applies to any other model that postulates that the expected (excess) return of a stock is a linear function of a certain characteristic. Though if the linearity of the postulated relation is taken for granted, one would suspect that an even more powerful test could be devised compared to the MR test. To make an analogy, consider studying the effect of years of schooling on wages in a regression model. If the form of the relation is not clear, a standard approach is to group years of schooling into a finite number of categories (0–5, 6–10, 11–15, etc.) and to include a dummy variable per category into the regression model. One then would have to show that the corresponding coefficients are strictly increasing to establish a strictly monotonic effect. On the other hand, if the relation is taken for granted as linear, a simpler (and more powerful) approach is to include years of schooling as a numerical regressor into the regression model and to then establish the corresponding slope as positive.

When the postulated relation is based on an empirical market anomaly rather than on economic theory, it seems necessary in general to choose a null hypothesis of no systematic relation. In this case the correct formulation of the testing problem is given by (9). ■

### 3 Designing tests for monotonicity in expected asset returns

It is assumed that estimators  $\hat{\Delta}_{T,i}$  for the individual expected return differentials  $\Delta_i$  are available. Here, the subscript  $T$  makes it explicit that the estimators depend on the given sample size  $T$ . The natural estimator for  $\Delta_i$  is simply the corresponding sample mean

$$\hat{\Delta}_{T,i} \equiv \bar{d}_{T,i} \equiv \frac{1}{T} \sum_{t=1}^T d_{t,i} . \quad (11)$$

However, the theory also allows for other choices  $\hat{\Delta}_{T,i}$ , such as robust estimators of means; for example, see [Scherer and Martin \(2005, Section 6.3\)](#).

Both expressions (5) and (10) initially suggest the following (overall) test statistic

$$\hat{\Delta}_T^{min} \equiv \min_i \hat{\Delta}_{T,i} . \quad (12)$$

However, as do P&T, we instead advocate the use of studentized individual test statistics. The main reason for studentization is to make sure that the individual test statistics are on the same scale, which can lead to large power gains; for example, see [Hansen \(2005\)](#) and [Romano and Wolf \(2005\)](#). The individual test statistic corresponding to  $\Delta_i$  is then given by

$$t_{T,i} \equiv \frac{\hat{\Delta}_{T,i}}{\hat{\sigma}_{T,i}}, \quad (13)$$

where  $\hat{\sigma}_{T,i}$  is a standard error for  $\hat{\Delta}_{T,i}$ . If the data constitute a time series, the standard error must be robust in that it must account for serial correlation; for example, one can use a HAC standard error as described in [Andrews \(1991\)](#) or a prewhitened HAC standard error as described in [Andrews and Monahan \(1992\)](#).

The following shall be assumed throughout, for  $i = 1, \dots, N$ : the estimator  $\hat{\Delta}_{T,i}$  is shift-equivariant and its standard error  $\hat{\sigma}_{T,i}$  is shift-invariant. That is, if the same constant  $c$  is added to all data points  $d_{1,i}, \dots, d_{T,i}$ , the estimator  $\hat{\Delta}_{T,i}$  is shifted by the same amount  $c$ , while its standard error  $\hat{\sigma}_{T,i}$  remains unchanged. This assumption holds true for any reasonable choice of estimator and standard error.

The (overall) test statistic is then defined as

$$t_T^{min} \equiv \min_i t_{T,i}. \quad (14)$$

Regardless of the choice of test statistic,  $\hat{\Delta}_T^{min}$  or  $t_T^{min}$ , one decides in favor of the alternative if the observed value of the test statistic is ‘sufficiently’ large. The difficulty lies in determining what constitutes ‘sufficiently’ large. That is, how does one determine a proper critical value for a level  $\alpha$  test? In the remainder of the paper, we shall restrict attention to the test statistic  $t_T^{min}$ . The issues are completely analogous for the alternative choice of test statistic  $\hat{\Delta}_T^{min}$ .

**Remark 3.1** All the tests for testing problem (9) that will be discussed have the following feature: a necessary condition for deciding in favor of  $H_1$  is that  $t_T^{min} > 0$ , which happens if and only if  $\min_i \hat{\Delta}_{T,i} > 0$ . This is a reasonable restriction accepted by the majority of the statistics community: to reject  $H_0$  in favor of the hypothesis  $H_1$  that the true parameter lies in a certain region, it is necessary that the observed parameter estimate lie in that region as well. For our purposes, to establish that  $\Delta \in R_3$  with any statistical significance, it is necessary that  $\hat{\Delta}_T \in R_3$  as well. Establishing that  $\Delta \in R_3$  when instead  $\hat{\Delta}_T \in R_1 \cup R_2$  would be considered ‘unreasonable’. Such ‘unreasonable’ tests do exist, but most statisticians shy away from them; for a discussion, see [Perlman and Wu \(1999\)](#).

An additional motivation for this feature is the concept of a monotone test.<sup>3</sup> In the context of testing problem (9), assume a test rejects  $H_0$  based on a vector of individual test statistics

---

<sup>3</sup>The term *monotone* here has a different meaning compared to the relation of expected returns being *monotonic*.

$\mathbf{t}_T \equiv (t_{T,1}, \dots, t_{T,N})'$ . Then the test is said to be *monotone* if it also rejects  $H_0$  based on any other vector of individual test statistics  $\tilde{\mathbf{t}}_T \equiv (\tilde{t}_{T,1}, \dots, \tilde{t}_{T,N})'$  which satisfies  $\tilde{\mathbf{t}}_T \geq \mathbf{t}_T$ . Such a requirement appears reasonable: if  $\tilde{\mathbf{t}}_T \geq \mathbf{t}_T$ , then  $\tilde{\mathbf{t}}_T$  should be considered at least as significant against  $H_0$  as  $\mathbf{t}_T$ . If the requirement of a monotone test is adopted, it can be shown that the feature of  $t_T^{\min} > 0$  being a necessary condition for rejection of  $H_0$  is rather innocuous; see the very end of Subsection 3.1.

On the other hand, the MR test of P&T for testing problem (7) can easily decide in favor of  $\Delta \in R_3$  when  $\hat{\Delta}_T \in R_2$ ; for example, see Subsection 6.1. This feature is due to the artifact that  $\Delta \in R_2$  is ruled out a priori by the assumptions of the MR test.

To put these concepts into perspective, consider an example where one test statistic  $t_{T,i}$  is small in magnitude and negative, say equal to  $-0.05$ , while all the other  $t_{T,j}$  are large in magnitude and positive, say equal to  $5.0$ . Since the corresponding  $\hat{\Delta}_T$  satisfies  $\hat{\Delta}_T \in R_2$ , none of the new tests for testing problem (9) to be discussed decide in favor of  $H_1$ . The interpretation would be that the relation is established as strictly monotonic ‘almost everywhere’ but just not everywhere. The fact that all the remaining  $t_{T,j}$  are highly significant cannot be used to establish that  $\Delta_i > 0$  as well. The MR test of P&T for testing problem (7) typically decides in favor of  $\Delta \in R_3$  in such a situation. Since the possibility of relation which is not systematic is ruled out on a priori grounds, it can be justified to consider a  $\hat{\Delta}_T$  that lies in  $R_2$  as evidence for  $\Delta \in R_3$ . As discussed in Perlman and Wu (1999), there also exist ‘unreasonable’ tests for testing problem (9) that can decide in favor of  $\Delta \in R_3$  in such a situation. However, such a test cannot have the desirable property of being a monotone test. ■

The general strategy to compute the critical value for a test against a one-sided (larger than) alternative at significance level  $\alpha$  is to derive the sampling distribution of the test statistic ‘under the null’ and to then take the  $1 - \alpha$  quantile of this derived null sampling distribution as the critical value.

Whether the testing problem is (7) or (9), the null hypothesis is *composite*, meaning that there are (infinitely) many parameters  $\Delta$  in the null space. Therefore, it is in principle not clear under which particular value of  $\Delta$  one should derive the sampling distribution ‘under the null’. One can deal with this dilemma in three ways: use the worst-case parameter under the null; try and guess the null parameter; or bound the true null parameter.

In the remainder of our discussion, we consider the data generating process (DGP) to be unknown but fixed, with the possible exception of  $\Delta$ . For example, when we talk about a worst-case distribution, as discussed below, we mean worst case with respect to  $\Delta$  only but not with respect to other features of the DGP, such as the correlation structure within the vector of return differentials  $\mathbf{d}_t$ , the various marginal distributions of the individual return differentials  $d_{t,i}$ , or the dependence structure of the  $\mathbf{d}_t$  over time. All such other features are



considered (unknown but) fixed.

More specifically, one can regard the unknown joint distribution of  $\hat{\Delta}_T$  as that of  $\mathbf{X} + \Delta$ , where  $\mathbf{X}$  is a random vector with a fixed distribution and  $\Delta$  is the unknown parameter. For example,  $\mathbf{X}$  could have the multivariate normal distribution with mean vector  $\mathbf{0}$  and (unknown but) fixed covariance matrix  $\Sigma$ ; or it can be multivariate- $t$ ; or something else. Regardless, under our maintained assumption that  $\hat{\Delta}_{T,i}$  is shift-equivariant while  $\hat{\sigma}_{T,i}$  is shift-invariant for all  $i = 1, \dots, N$ , it then follows that larger values of  $\Delta$  lead to larger values of the test statistic  $t_T^{min}$ . In particular, if  $\Delta_1 \leq \Delta_2$ , then  $\mathbb{P}_{\Delta_1}\{t_T^{min} > c\} \leq \mathbb{P}_{\Delta_2}\{t_T^{min} > c\}$ , for any real number  $c$ .

### 3.1 Worst-case approaches

The worst case under the null corresponds to the parameter  $\Delta_0$  in the null space that results in the largest critical value possible. Specifically, if we consider the  $1 - \alpha$  quantile of the distribution of a given test statistic under  $\Delta_0$ , then the most conservative approach is to take the largest such critical value as  $\Delta_0$  ranges over the null hypothesis parameter space. Intuitively, this specific parameter is the parameter in the null space that is closest to the alternative space. In other words, it lies on the boundary between the null and the alternative space. In statistical lingo, it gives rise to the *least favorable distribution*, where the term ‘favorable’ is to be understood with respect to the ability to reject false hypotheses: the larger the critical value of the test, the smaller is its power against any given alternative.

The motivation here is quite simple: if one uses the worst-case parameter under the null to derive the critical value, then test will work for any parameter under the null in the sense that rejection probability will be at most  $\alpha$ , but can be less. This is because, by definition, the critical value for any other parameter in the null space is at least as large as the critical value derived from the worst-case parameter.

Crucially, the worst-case parameter under the null depends on the choice of testing problem: (7) or (9). For testing problem (7), it is easy to see that the worst-case parameter is given by

$$\Delta_0^{\text{WC},(7)} \equiv \mathbf{0} \equiv (0, 0, \dots, 0)' . \quad (15)$$

The resulting test is called the MR test (for monotonic relation test) by P&T.

But this is no longer true for testing problem (9). Consider the special case  $N = 2$ . For any positive number  $\delta$ , both the parameters  $(\delta, 0)$  and  $(0, \delta)$  will lead to a larger critical value compared to the parameter  $(0, 0)$ . (However, which of the two leads to the largest critical value is not necessarily clear.) Furthermore, increasing  $\delta$  will increase the two critical values. Taking this logic to the limit, one realizes that the worst-case parameter is one of the following two:  $(\infty, 0)$  or  $(0, \infty)$ .

**Remark 3.2** Any parameter  $\Delta$  can only have finite entries. What we mean by the worst-case parameter being equal to  $(\infty, 0)$ , say, is the following. The critical value under the parameter  $(\infty, 0)$  is obtained as the limit of the critical value under  $(\delta, 0)$ , as  $\delta$  tends to infinity. For all practical purposes, the critical value can be obtained under  $(\delta^{\text{Big}}, 0)$  where  $\delta^{\text{Big}}$  is a very large number, such as  $\delta^{\text{Big}} = 10^6$ .

Furthermore, the worst-case parameter is not necessarily unique. For example, the two choices  $(\infty, 0)$  and  $(0, \infty)$  can lead to the same critical value. In such a case, ‘the’ worst-case parameter is taken to be an arbitrary element of the set of all worst-case parameters. ■

For a general dimension  $N$ , the worst-case parameter for testing problem (9) is of the form

$$\Delta_0^{\text{WC},(9)} \equiv (\infty, \dots, \infty, 0, \infty, \dots, \infty)' . \quad (16)$$

That is, one of the entries is zero and all the other entries are infinity. In case it is not clear a priori which particular position of the zero entry leads to the largest possible critical value, one has to try out all possible positions in principle. However, as will be seen, under mild regularity conditions, all such least favorable points lead to an asymptotic critical value of  $z_{1-\alpha}$ , the  $1 - \alpha$  quantile of the standard normal distribution.

We call the resulting test the Cons test, short for conservative test. A connection to the generalized likelihood ratio test for testing problem (9) is given in Appendix A.

Recall the concept of a monotone test from Remark 3.1. It can be shown that for testing problem (9), in an approximate set-up, the Cons test is actually uniformly most powerful (UMP) among all monotone level  $\alpha$  tests; see Appendix B. As the critical value of the Cons test is positive, this provides another justification to restrict attention to tests which require  $t_T^{\min} > 0$  as a necessary condition to reject  $H_0$  in testing problem (9).

### 3.2 Constrained estimation of the null parameter

By design, using the worst-case parameter to derive the critical value of a test leads to conservative inference in general. That is, the rejection probability under an arbitrary parameter in the null hypothesis parameter space will often be less than the significance level of the test.

Take the example of testing problem (9) with  $N = 2$ . If the worst-case parameter is  $(\infty, 0)$  but the true parameter is  $(1, 0)$ , say, then the rejection probability will be strictly smaller than  $\alpha$  (at least for any ‘reasonable’ DGP). This is because the critical value under  $(1, 0)$  is strictly smaller than the critical value under  $(\infty, 0)$ .

Therefore, it would be better to compute the critical value under the true parameter  $(1, 0)$ . The dilemma is that the true parameter is not known. It is then tempting to estimate the true parameter from the data and to use a corresponding parameter to compute the critical value

rather than to use the worst-case parameter. In doing so, the estimate must be constrained to lie in the null parameter space, of course.

One observes  $\hat{\Delta}_T \equiv (\hat{\Delta}_{T,1}, \dots, \hat{\Delta}_{T,N})'$ . As discussed before in Remark 3.1, we only have to consider the case of  $\hat{\Delta}_T^{min} > 0$ , since any ‘reasonable’ test for testing problem (9) will not reject if  $\hat{\Delta}_T^{min} \leq 0$ . A constrained estimate for the null parameter is then of the form

$$\Delta_0^{Est} \equiv (\hat{\Delta}_{T,1}, \dots, \hat{\Delta}_{T,i-1}, 0, \hat{\Delta}_{T,i+1}, \dots, \hat{\Delta}_{T,N})' . \quad (17)$$

It is not clear a priori which position for the zero entry will lead to the largest critical value, so in principle one has to try out all  $N$  possibilities. Since this approach is based on the observed estimate  $\hat{\Delta}_T$  rather than the true parameter  $\Delta$ , one cannot guarantee that the resulting test will work (in the sense of ensuring rejection probability of at most  $\alpha$  for any null parameter) for finite samples.<sup>4</sup>

We will call the resulting test the Constrained-Estimate (CE) test.

### 3.3 Constrained bounding of the null parameter

A more conservative approach first constructs an upper joint confidence region for  $\Delta$  and then extracts the null parameter from the corresponding upper limits rather than from the point estimate  $\hat{\Delta}_T$ . Crucially, to ensure that the resulting test works well in finite samples, one has to downward adjust the significance level of the test depending on the confidence level used for the joint confidence region.

More specifically, consider a joint confidence region for  $\Delta$  at confidence level  $1 - \beta$  of the form

$$(-\infty, \hat{u}_{T,1}] \times (-\infty, \hat{u}_{T,2}] \times \dots \times (-\infty, \hat{u}_{T,N}] . \quad (18)$$

For example, the vector of upper limits  $\hat{u}_T \equiv (\hat{u}_{T,1}, \hat{u}_{T,2}, \dots, \hat{u}_{T,N})'$  can be computed by the single-step method described in Romano and Wolf (2005); for completeness, this method is briefly summarized in Appendix C. It satisfies  $\hat{u}_{T,i} > \hat{\Delta}_{T,i}$ , for all  $i$ . The constrained bounded parameter then lies on the intersection of the region (18) together with the null hypothesis parameter space  $R_1 \cup R_2$  and, therefore, is of the form

$$\Delta_0^{Bound} \equiv (\hat{u}_{T,1}, \dots, \hat{u}_{T,i-1}, 0, \hat{u}_{T,i+1}, \dots, \hat{u}_{T,N})' . \quad (19)$$

Again, it is not clear a priori which position for the zero entry will lead to the largest critical value, so in principle one has to try out all  $N$  possibilities. Crucially, the critical value is now computed as the  $1 - \alpha + \beta$  quantile of the corresponding null sampling distribution of  $t_T^{min}$  to ensure the overall validity of the test in finite samples. The intuition here is that with

---

<sup>4</sup>Since  $\hat{\Delta}_T$  converges to  $\Delta$  in probability as  $T$  goes to infinity, the approach works asymptotically, though; see Theorem 4.1.

probability of (at most)  $\beta$  the confidence region (18) will not contain the true parameter  $\Delta$ ; and one must then adjust for this fact by decreasing the nominal significance level of the test from  $\alpha$  to  $\alpha - \beta$ . In particular, one must choose  $\beta < \alpha$  in the computation of the confidence region (18).

We will call the resulting test the Two-Step test, due to its two-step nature. Related two-step inference procedures have been previously suggested by Loh (1985), Romano and Wolf (2000), and Hansen (2003), among others.

## 4 Implementing tests for monotonicity in expected asset returns

Having chosen a specific null parameter  $\Delta_0$ , by any one of the forms (15), (16), (17), or (19), one is left to derive or approximate the implied sampling distribution of the test statistic  $t_T^{min}$ , and to then take the appropriate quantile as the critical value of the test. In general, one takes the  $1 - \alpha$  quantile; with the exception of the Two-Step test, where one takes the  $1 - \alpha + \beta$  quantile instead.

It is not possible to derive the implied sampling distribution analytically. As a feasible solution, the distribution can be approximated via the bootstrap. To this end, create a bootstrap data set  $\{\mathbf{r}_1^*, \mathbf{r}_2^*, \dots, \mathbf{r}_T^*\}$  by resampling from the observed data  $\{\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_T\}$ . Denote the estimator of  $\Delta_i$  computed from the bootstrap data by  $\hat{\Delta}_{T,i}^*$  and let  $\Delta_T^* \equiv (\hat{\Delta}_{T,1}^*, \dots, \hat{\Delta}_{T,N}^*)'$ . Similarly, denote the standard error for  $\hat{\Delta}_{T,i}^*$  computed from the bootstrap data by  $\hat{\sigma}_{T,i}^*$ . The following algorithm details how the critical value of the test is approximated via the bootstrap.

### Algorithm 4.1 (Computation of the critical value)

- Generate bootstrap data  $\{\mathbf{r}_1^*, \mathbf{r}_2^*, \dots, \mathbf{r}_T^*\}$  and compute statistics  $\hat{\Delta}_{T,i}^*$  and  $\hat{\sigma}_{T,i}^*$  from these data, for  $i = 1, \dots, N$
- Compute  $t_{T,i}^*$ , for  $i = 1, \dots, N$ , defined as:

$$t_{T,i}^* \equiv \frac{\hat{\Delta}_{T,i}^* - \hat{\Delta}_{T,i} + \Delta_{0,i}}{\hat{\sigma}_{T,i}^*} \quad (20)$$

- Compute  $t_T^{min,*} \equiv \min_i t_{T,i}^*$
- Repeat this process  $B$  times, resulting in statistics  $t_{T,1}^{min,*}, \dots, t_{T,B}^{min,*}$
- The empirical distribution of these statistics  $t_{T,1}^{min,*}, \dots, t_{T,B}^{min,*}$  is the bootstrap approximation to the sampling distribution of  $t_T^{min}$  under the parameter  $\Delta_0$
- In particular, the corresponding  $1 - \alpha$  empirical quantile serves as the critical value of the test; with the exception of the Two-Step test, where one takes the corresponding  $1 - \alpha + \beta$  empirical quantile instead

**Remark 4.1** Which particular bootstrap method should be used to generate the bootstrap data  $\{\mathbf{r}_1^*, \mathbf{r}_2^*, \dots, \mathbf{r}_T^*\}$  from the observed data  $\{\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_T\}$  depends on the underlying DGP. If the original data are i.i.d., one can use the standard bootstrap of [Efron \(1979\)](#). If the original data constitute a time series, one needs to employ a proper time series bootstrap and several choices are available; for example, see [Lahiri \(2003\)](#). [Romano and Wolf \(2005\)](#) suggest the circular block bootstrap of [Politis and Romano \(1992\)](#) while P&T suggest the stationary bootstrap of [Politis and Romano \(1994\)](#).

Either way, it is important to point out that one simply ‘resamples from the observed data’ as opposed to ‘resampling from some null-enforced data’, where  $\Delta_0$  is the true parameter. This is because one adds the number  $(\Delta_{0,i} - \hat{\Delta}_{T,i})$  to the bootstrap estimate  $\hat{\Delta}_{T,i}^*$  in the numerator of the bootstrap test statistic (20) to approximate the sampling distribution of  $t_T^{min}$  under  $\Delta_0$  in the end, as desired. This approach is (asymptotically) equivalent to instead ‘resampling from some null-enforced data’ and to then computing the bootstrap test statistic as

$$t_{T,i}^* \equiv \frac{\hat{\Delta}_{T,i}^*}{\hat{\sigma}_{T,i}^*}. \quad (21)$$

The latter approach is more complicated and could be carried out by some empirical likelihood method; for example, see [Owen \(2001\)](#). ■

Of course, so far Algorithm 4.1 is merely a ‘recipe’. Certain conditions are needed to ensure that the resulting critical value obtained by the bootstrap is a valid approximation. To this end, we will make use of the following high-level assumptions.

- (A1)  $\sqrt{T}(\hat{\Delta} - \Delta)$  converges in distribution to  $N(\mathbf{0}, \Omega)$ , for some positive definite  $N \times N$  matrix  $\Omega$  with typical element  $\omega_{i,j}$ .
- (A2) The bootstrap consistently estimates this limiting distribution, namely  $\sqrt{T}(\hat{\Delta}^* - \hat{\Delta})$  converges in distribution to  $N(\mathbf{0}, \Omega)$  in probability.
- (A3) Both  $\sqrt{T}\hat{\sigma}_{T,i}$  and  $\sqrt{T}\hat{\sigma}_{T,i}^*$  converge in probability to  $\sqrt{\omega_{i,i}}$ , for  $i = 1, \dots, N$ .

These high-level assumptions are rather weak and have been verified for many applications of interest in the statistics literature.

One can then show the following theorem, the proof of which can be found in Appendix D:

**Theorem 4.1** *Assume (A1)–(A3) and compute the critical value of the tests as detailed in Algorithm 4.1. Then, as  $T \rightarrow \infty$ :*

- (ia) *For any  $\Delta \in R_1$  and for any  $\Delta \in R_2$  with  $\min_i \Delta_i < 0$ , the limiting rejection probability of the MR test is no larger than  $\alpha$ .*

- (ib) For any  $\Delta$  on the boundary of  $R_2$ , excluding the origin, the limiting rejection probability of the MR test is strictly larger than  $\alpha$ .
- (ic) For any  $\Delta \in R_3$ , the limiting rejection probability of the MR test is one.
- (iia) For any  $\Delta \in R_1 \cup R_2$ , the limiting rejection probabilities of the Cons, CE, and Two-Step tests are no larger than  $\alpha$ .
- (iib) For any  $\Delta \in R_3$ , the limiting rejection probabilities of the Cons, CE, and Two-Step tests are one.

**Remark 4.2** An important implication for applied work stems from part (ib) of the theorem. If  $\Delta$  lies on the boundary of  $R_2$ , excluding the origin (that is, if  $\min_i \Delta_i = 0$  but  $\max_i \Delta_i > 0$ ), then the limiting rejection probability of the test is strictly larger than  $\alpha$ . In the special case where the test statistics are asymptotically uncorrelated, the probability of a type 1 error can, approximately, be as large as  $\alpha^{1/N}$ ; see Remark B.1 in Appendix B. So even when  $\Delta \in R_2$ , the test can decide in favor of  $\Delta \in R_3$  with quite high probability.

Actually, this is not to be considered a failure of the MR test. As pointed out before, the MR test is specifically designed for the case where  $\Delta$  can a priori only lie either in  $R_1$  or in  $R_3$ . Instead, it is rather the failure of the applied researcher to apply the MR test to settings for which it was not designed, namely to settings for which  $\Delta$  can a priori also lie in  $R_2$ . To interpret the rejection of the null hypothesis for the MR as evidence for  $\Delta \in R_3$  is then not necessarily justified. ■

**Remark 4.3** One might feel encouraged by part (ia) of the theorem to trust the MR test also for  $\Delta$  ‘within’  $R_2$ , that is, for  $\Delta \in R_2$  with  $\min \Delta_i < 0$ . Indeed, the result guarantees that the limiting rejection probability will be at most  $\alpha$  in this case. However, the rejection probability in finite samples can actually be well above  $\alpha$ ; see Section 5.

Similarly, the CE test can display rejection probabilities above  $\alpha$  in finite samples even if  $\Delta \in R_1 \cup R_2$ . On the other hand, the Cons test and the Two-Step test appear to always successfully control the probability of a type 1 error; see Section 5.

This shows the importance of considering both the asymptotic properties and the finite-sample properties of a testing procedure before applying it to real data. ■

**Remark 4.4** The worst-case parameter  $\Delta_0^{\text{WC},(9)}$  for the Cons test is of the form (16), that is, all entries are infinity apart from a single zero entry. But then it is easy to see that under Assumptions (A1)–(A3), the limiting critical value of the test is simply  $z_{1-\alpha}$ , the  $1 - \alpha$  quantile of the standard normal distribution; regardless of the position of the zero entry. Since the bootstrap critical value is valid asymptotically only in the first place, one can simply use  $z_{1-\alpha}$  as the critical value of the test, foregoing any application of the bootstrap. Indeed, the

two approaches are equivalent to first order, since the bootstrap critical value will converge to  $z_{1-\alpha}$  in probability. However, it should be pointed out that the computation of the critical value via the bootstrap typically results in better finite-sample properties of the test, as the bootstrap is better able to capture skewness; for example, see [Hall \(1992\)](#). ■

## 5 Monte Carlo study

This section examines the finite-sample properties of the various tests via Monte Carlo studies; both in terms of controlling the probability of a type 1 error under the null and in terms of power.

The following tests are included in the study:

- **(MR)** The MR test of P&T described in Subsection [3.1](#)
- **(CE)** The CE test described in Subsection [3.2](#)
- **(TwoStep)** The Two-Step test described in Subsection [3.3](#)
- **(Cons)** The Cons test described in Subsection [3.1](#)

We use  $\alpha = 0.05$  as the significance level and  $\beta = 0.01$  for the Two-Step test. The sample size is  $T = 120$  always. All empirical rejection probabilities are based on 20,000 simulations. The number of bootstrap repetitions is  $B = 499$ . Return differential vectors  $\mathbf{d}_t$  are generated in i.i.d. fashion and, accordingly, the standard bootstrap of [Efron \(1979\)](#) is employed.

The individual test statistics  $t_{T,i}$  are computed as follows

$$t_{T,i} \equiv \frac{\hat{\Delta}_{T,i}}{\hat{\sigma}_{T,i}} \quad \text{where} \quad \hat{\Delta}_{T,i} \equiv \bar{d}_{T,i} \equiv \frac{1}{T} \sum_{t=1}^T d_{t,i} \quad \text{and} \quad \hat{\sigma}_{T,i}^2 \equiv \frac{1}{T-1} \sum_{t=1}^T (d_{t,i} - \bar{d}_{T,i})^2. \quad (22)$$

### 5.1 Null behavior for $N = 2$

We start with the extreme case of  $N = 2$  to more easily judge the effect of correlation within a return differential vector. The return differentials are generated as

$$\mathbf{d}_t = \begin{pmatrix} d_{t,1} \\ d_{t,2} \end{pmatrix} \sim N \left( \begin{pmatrix} \Delta_1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right), \quad (23)$$

with  $\Delta_1 \in \{0, 0.025, 0.05, \dots, 0.45, 0.475, 0.5\}$  and  $\rho \in \{-0.5, 0.0, 0.5\}$ .

The parameter  $\mathbf{\Delta}$  lies on the boundary of region  $R_2$  always. For the special case of  $\Delta_1 = 0$ , it also lies on the boundary of region  $R_1$ . However, it never lies in region  $R_3$ , that is, the alternative of strict monotonicity is never true.

The empirical rejection probabilities (ERPs) of the various tests are presented in [Fig. 1](#).

One can see that MR successfully controls the probability of a type 1 error for  $\Delta_1 = 0$  but is otherwise liberal. This problem increases as the correlation  $\rho$  decreases.

By design, the other three tests are conservative at  $\Delta_1 = 0$ , with their rejection probability increasing with  $\Delta_1$ .

CE can also be liberal for  $\rho \in \{-0.5, 0.0\}$ , though much less compared to MR.

Both TwoStep and Cons are successful in controlling the probability of a type 1 error across all values of  $\Delta_1$  and  $\rho$ , with Cons approaching the nominal level  $\alpha$  slightly faster.

## 5.2 Null behavior for $N = 10$

We continue with the more relevant case of  $N = 10$  and consider two types of dependence structure for the return differentials. Specifically, the return differentials are generated as

$$d_t \sim N(\mathbf{\Delta}, \mathbf{\Omega}) ,$$

where  $\mathbf{\Delta} = (\Delta_1, \dots, \Delta_{10})'$  and  $\mathbf{\Omega}$  is either the identity covariance matrix of dimension 10 or a Toeplitz covariance matrix of dimension 10 with typical element  $\omega_{i,j} = 0.9^{|i-j|}$ . In the former case, the return differentials of a common period  $t$  are mutually independent; in the latter case, there is strong dependence for return differentials close to each other and weak dependence for return differentials far apart from each other.

We consider three basic designs for  $\mathbf{\Delta}$ :

- **(D1)**  $\mathbf{\Delta} = (\Delta, \dots, \Delta, -\Delta/10, \dots, -\Delta/10)'$
- **(D2)**  $\mathbf{\Delta} = (\Delta, \dots, \Delta, -\Delta)'$
- **(D3)**  $\mathbf{\Delta} = (\Delta, \dots, \Delta, 0)'$

with  $\Delta \in \{0, 0.025, 0.05, \dots, 0.45, 0.475, 0.5\}$ . In particular, in

- **(D1)** the last five elements are generally negative, though of a smaller magnitude.
- **(D2)** the last element is generally negative, and of the same magnitude.
- **(D3)** the last element is always zero.

When  $\Delta = 0$ , then  $\mathbf{\Delta}$  lies on the boundary of regions  $R_1$  and  $R_2$  for all three designs. When  $\Delta > 0$ , then: for design 1 and design 2,  $\mathbf{\Delta}$  lies within region  $R_2$ ; for design 3,  $\mathbf{\Delta}$  lies on the boundary of region  $R_2$ . However, it never lies in region  $R_3$ , that is, the alternative of strict monotonicity is never true.

For an illustration in terms of the corresponding vectors of expected returns, which have dimension  $N + 1 = 11$ , see Fig. 2; the value  $\Delta = 0.1$  is used there.



The empirical rejections probabilities (ERPs) of the various tests are presented in Fig. 3 and Fig. 4.

One can see that MR is often liberal for all three designs. In particular, for design 3 it is always liberal except for  $\Delta = 0$ . And for large values of  $\Delta$ , the probability of a type 1 error rises as high as about 70% for the independent case. But even in design 1, where half of the expected return differentials are actually negative, albeit at a smaller magnitude, the test is generally liberal and the probability of a type 1 error can be as high as 15%. The overrejection problem is less pronounced for the dependent case (that is, for the Toeplitz covariance matrix) but still noticeable.

CE is somewhat liberal in design 3 for the independent case, though not for the dependent case and also never in design 1 or design 2.

The performances of TwoStep and Cons are virtually indistinguishable. In design 3, the probability of a type 1 error tends to 0.05 as  $\Delta$  increases; it never lies above 0.05. In design 1 and design 2, the probability of a type 1 error is rather flat near zero. This holds true both for the independent and the dependent case. Note here that since  $\Delta$  lies within region  $R_2$  in design 1 and design 2 rather than on the boundary, the probability of a type 1 error cannot approach 0.05 for CE, TwoStep, and Cons.

### 5.3 Alternative behavior for $N = 10$ and $N = 5$

We now turn to performance with respect to power instead.

Ceteris paribus, it is reasonable to expect that power increases when the number of return differentials,  $N$ , is decreased. If the applied researcher uses a very large number  $N$ , then it will be rather (more) difficult for her to establish a positive expected return differential for each category.

To examine this effect, we consider the case where the expected return differentials are all the same and employ  $N = 10$  or  $N = 5$ . To keep everything else equal, the expected return differentials must be twice as large for  $N = 5$  compared to  $N = 10$ . This results in what we call designs 4 and 5:

- **(D4)**  $\Delta = (\Delta, \dots, \Delta)'$  with  $\Delta \in \{0, 0.025, 0.05, \dots, 0.45, 0.475, 0.5\}$ .
- **(D5)**  $\Delta = (\Delta, \Delta, \Delta, \Delta, \Delta)'$  with  $\Delta \in \{0, 0.05, 0.1, \dots, 0.95, 1.0\}$ .

For example,  $\Delta = 0.2$  in design 4 with  $N = 10$  corresponds to  $\Delta = 0.4$  in design 5 with  $N = 5$ . In other words, by reducing the number of categories by half, the expected return differentials must get doubled.

Also for  $N = 5$  we consider two cases for the covariance matrix  $\Omega$  of the return differentials: the identity covariance matrix of dimension 5 or a Toeplitz covariance matrix of dimension 5 with typical element  $\omega_{i,j} = 0.8^{|i-j|}$ .

In both designs 4 and 5, the parameter  $\Delta$  lies on the boundary of regions  $R_1$  and  $R_2$  when  $\Delta = 0$  and in region  $R_3$  when  $\Delta > 0$ . So the alternative of strict monotonicity is generally true, with the exception of  $\Delta = 0$ .

The empirical rejections probabilities (ERPs) of the various tests are presented in Fig. 5 and Fig. 6. Naturally, MR has the highest power, followed by CE, followed by Cons and TwoStep.

One can see that for all tests the power increases when  $N$  is reduced from  $N = 10$  to  $N = 5$ , but on relative terms the most for TwoStep and Cons. For the independent case, take the combination  $(N = 10, \Delta = 0.2)$  in design 4 that corresponds, one-to-one, to the combination  $(N = 5, \Delta = 0.4)$  in design 5. While before the power of those two tests in design 3 was slightly below 0.05, it has now risen to almost one in design 4.

It is also noteworthy to point out that the differences in power are generally much reduced for the case of dependent returns (that is, when the covariance matrix of the return differentials is a Toeplitz matrix); see Fig. 6 versus Fig. 5.

In applied work, researchers often construct either  $N + 1 = 10$  or  $N + 1 = 5$  categories based on portfolio sorts using either deciles or quintiles, respectively. For testing monotonicity of expected asset returns, it would then be preferable to use quintiles instead of deciles in terms of the power of the tests. On the other hand, if monotonicity can be established for a larger value of  $N$ , this can be considered more convincing evidence. In the end, if the choice of  $N$  is not predetermined, then it is up to the applied researcher to use her judgment in selecting a suitable number.

## 5.4 Overall recommendations

When the parameter space for  $\Delta$  is  $\mathbb{R}^N$ , the MR test is not recommended. The test can decide in favor of  $\Delta \in R_3$  with probability much exceeding the nominal significance level when in fact  $\Delta \notin R_3$ . This especially happens when  $\Delta$  lies on the boundary of  $R_2$ . However, it can also happen even when  $\Delta$  lies within  $R_2$ .

The CE test has related problems, though to a much smaller extent. Still, we would not recommend it for general use.

The only two tests that are safe to use in general are the Two-Step and the Cons tests. The performance of these two tests is rather similar. Since the Cons test is much easier to implement, it is the one we recommend for practical use.

The situation is different when it is known a priori that  $\Delta \in R_1 \cup R_3$ . In this case, the MR test successfully controls the probability of a type 1 error. And since the MR test has the highest power of all four tests, it is then the preferred one.

Regardless, in terms of the power of the tests it is advisable to keep the number of return

categories  $N + 1$  to a ‘reasonable’ minimum. For example, basing portfolio sorts on quintiles rather than on deciles will result in higher power of the various tests.

## 6 Empirical applications

In this section, we revisit two of the empirical applications of P&T before considering some additional new ones.

### 6.1 Revisiting two empirical applications of P&T

P&T present two main empirical applications: one application concerning equity returns based on portfolios formed on CAPM beta and another application concerning bond term premia based on maturity. Their two corresponding plots of sample average returns and sample average term premia, respectively, according to return categories are reproduced in Fig. 7.

We first turn to the application concerning equity returns. As detailed in their Subsection 4.1, the MR test of P&T decides in favor of a monotonic relation, that is, in favor of the alternative  $\Delta \in R_3$ . In contrast, since it is clear from the upper half of Fig. 7 that  $\hat{\Delta}_T \in R_2$ , none of the new tests presented in this paper decide in favor of  $\Delta \in R_3$ ; see Remark 3.1. Instead, all these tests ‘accept’ the null hypothesis of  $\Delta \in R_1 \cup R_2$ . In view of these contradictory results, the question then becomes whether the linearity of the relation between the CAPM beta of a stock and its expected return is taken for granted or not; see Remark 2.1.

We next turn to the application concerning bond term premia. As detailed in their Subsection 4.3, the MR test of P&T does not decide in favor of a monotonic relation, that is, in favor of the alternative  $\Delta \in R_3$ . Since it is clear from the lower half of Fig. 7 that  $\hat{\Delta}_T \in R_2$ , the new tests presented in this paper all agree with this conclusion.

### 6.2 Additional empirical applications

We present three additional empirical applications where we consider the effect of short-term reversal; here, short-term reversal is one-month reversal. We do this separately for small-size firms, mid-size firms, and large-size firms. To this end, we use a data set provided by Kenneth R. French labeled “25 Portfolios Formed on Size and Short-Term Reversal (5 x 5)”.<sup>5</sup> In particular, we consider monthly average value-weighted returns from 01/1927 until 12/2010. In total, there are 25 portfolios. We use the first five, which correspond to quintile portfolios formed on short-term reversal for small-size firms, the middle five, which correspond to quintile portfolios formed on short-term reversal for mid-size firms, and the last five, which correspond

---

<sup>5</sup>Data available at <http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data-library.html>.

to quintile portfolios formed on short-term reversal for large-size firms. The sample size is  $T = 1,008$ .

The test statistics are obtained as

$$t_{T,i} \equiv \frac{\hat{\Delta}_{T,i}}{\hat{\sigma}_{T,i}} \quad \text{where} \quad \hat{\Delta}_{T,i} \equiv \bar{d}_{T,i} \equiv \frac{1}{T} \sum_{t=1}^T d_{t,i} \quad (24)$$

and the standard errors  $\hat{\sigma}_{T,i}$  are computed via a kernel method using the quadratic-spectral kernel and the automatic choice of bandwidth proposed by [Andrews \(1991\)](#). The critical values of the tests are computed via Algorithm 4.1. We employ the circular block bootstrap of [Politis and Romano \(1992\)](#) with a block size of  $b = 12$ , based 10,000 bootstrap resamples always. The choice  $b = 12$  seems justified as none of the data display strong serial correlation.<sup>6</sup>

The average returns on quintile portfolios formed on short-term reversal are displayed in Fig. 8. For all three firm sizes, the observed relation is strictly monotonic decreasing. However, the ‘steepness’ of the relation diminishes in magnitude as the firm size goes up.

Since the observed relations are strictly monotonic decreasing rather than increasing, all return differentials were multiplied by negative one before feeding the data to the testing procedures. The results of the MR and Cons tests, for a nominal level of  $\alpha = 0.05$ , are presented in Table 1. Both tests establish a strictly monotonic relation for small-size firms and fail to do so for large size-size firms. However, the test results differ for mid-size firms: the MR test establishes a strictly monotonic relation while the Cons test does not.

## 7 Conclusions

In many instances, empirical research in finance seeks to address whether a strictly monotonic increasing relation exists between an asset’s expected return and some underlying characteristic of interest. For example, in the context of equity returns, such a characteristic might be CAPM beta, size, book-to-market, momentum, or reversal; in the context of bond returns, such a characteristic might be maturity or rating quality. If this characteristic is ordered into more than two categories, then a convincing test needs to establish monotonicity over all categories. In other words, a simple test of “high minus low”, only comparing the two most extreme categories, is not convincing.

In a recent paper, [Patton and Timmermann \(2010\)](#) propose a general test, taking all categories of the underlying characteristic into account. Compared to previous related proposals, they are, to our knowledge, the first ones to postulate a strictly increasing monotonic relation as the alternative hypothesis of test, rather than as the null hypothesis. This is the correct

---

<sup>6</sup>The qualitative results, in terms of rejecting the null hypothesis or not, remain unchanged if a block size of  $b = 24$  is used instead.

formulation if, indeed, the goal is to establish strict monotonicity with a quantifiable statistical significance.

On the other hand, they postulate a weakly decreasing monotonic relation as the null hypothesis. Compared to allowing, more generally, for a non-systematic relation under null, this approach results in a smaller critical value of the test and thereby in higher power. Unfortunately, if a non-systematic relation is actually possible in case the alternative is not true, this test does not successfully control the probability of a type 1 error (that is, of falsely rejecting the null). As a result, it can become quite likely to ‘establish’ a strictly monotonic increasing relation when in reality it does not exist.

In this paper, we have proposed some alternative tests that allow for a non-systematic relation under the null and successfully control the probability of a type 1 error. As is unavoidable, such tests will always have lower power compared to the test of [Patton and Timmermann \(2010\)](#). Moreover, a necessary condition for such tests to decide in favor of the alternative is that a strictly monotonic increasing relations is observed in the data in the first place. It appears that this criterion can be difficult to meet already for many relations of practical interest. A potential reason is the large amount of noise often contained in financial data. And even if strict monotonicity is actually observed in the data, the new tests we have proposed still do not always decide in favor of the alternative (as can happen when the observed differences between categories are small relative to the noise contained in the data). Therefore, in general, it might actually be a rather ambitious goal trying to establish strict monotonicity of expected financial returns.

The goal becomes more realistic when a non-systematic relation can be ruled out on a priori grounds, that is, when the null hypothesis can be restricted to a weakly monotonic decreasing relation. In this case, the test of [Patton and Timmermann \(2010\)](#) successfully controls the probability of a type 1 error and exhibits attractive power properties.

## A Connection to the quasi likelihood ratio test

Suppose  $\mathbf{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Omega})$ , with  $\boldsymbol{\Omega}$  a known positive definite matrix. For testing problem (9) applied to the vector  $\boldsymbol{\mu}$ , namely

$$H_0 : \boldsymbol{\mu} \in R_1 \cup R_2 \quad \text{vs.} \quad H_1 : \boldsymbol{\mu} \in R_3 , \quad (25)$$

the generalized likelihood ratio (QLR) test rejects for large values of

$$\frac{\sup_{\boldsymbol{\mu} \in \mathbb{R}^N} \exp \left[ -\frac{1}{2} (\mathbf{X} - \boldsymbol{\mu})' \boldsymbol{\Omega}^{-1} (\mathbf{X} - \boldsymbol{\mu}) \right]}{\sup_{\boldsymbol{\mu} \in R_3^c} \exp \left[ -\frac{1}{2} (\mathbf{X} - \boldsymbol{\mu})' \boldsymbol{\Omega}^{-1} (\mathbf{X} - \boldsymbol{\mu}) \right]} , \quad (26)$$

where  $R_3^c = R_1 \cup R_2$ . Assume that  $\mathbf{X} > \mathbf{0}$ , or the test accepts  $H_0$  anyway. Since the supremum in the numerator occurs when  $\boldsymbol{\mu} = \mathbf{X}$ , the QLR test rejects for large values of

$$\inf_{\boldsymbol{\mu} \in R_3^c} (\mathbf{X} - \boldsymbol{\mu})' \boldsymbol{\Omega}^{-1} (\mathbf{X} - \boldsymbol{\mu}) . \quad (27)$$

In the case where  $\boldsymbol{\Omega}$  is a diagonal matrix, the infimum occurs at some  $\boldsymbol{\mu}$  that satisfies

$$\mu_i = 0 \text{ for some } i \in \{1, \dots, N\} \quad \text{and} \quad \mu_j = X_j \text{ for all } j \in \{1, \dots, N\} \text{ with } j \neq i . \quad (28)$$

The resulting value of (27) is then  $\min_i X_i^2 / \omega_{i,i}$ , and rejecting for large values is equivalent to rejecting for large  $\min_i X_i / \sqrt{\omega_{i,i}}$ .

In case  $\boldsymbol{\Omega}$  is not known, but a suitable estimator  $\hat{\boldsymbol{\Omega}} = \text{diag}(\hat{\omega}_{1,1}, \dots, \hat{\omega}_{N,N})$  is available, a feasible version of the QLR test rejects for large values of  $\min_i X_i / \sqrt{\hat{\omega}_{i,i}}$ .

In the set-up of the paper, Assumptions (A1)–(A3) imply the following approximate sampling distribution, at least for large sample sizes:

$$\hat{\boldsymbol{\Delta}}_T \dot{\sim} N(\boldsymbol{\Delta}, \boldsymbol{\Omega}/T) \quad \text{or, equivalently,} \quad \sqrt{T} \hat{\boldsymbol{\Delta}}_T \dot{\sim} N(\sqrt{T} \boldsymbol{\Delta}, \boldsymbol{\Omega}) , \quad (29)$$

where the symbol  $\dot{\sim}$  denotes “is approximately distributed as”.

Hence — with  $\sqrt{T} \hat{\boldsymbol{\Delta}}_T$  playing the role of  $\mathbf{X}$ , with  $\sqrt{T} \boldsymbol{\Delta}$  playing the role of  $\boldsymbol{\mu}$ , and with  $\sqrt{T} \hat{\sigma}_{T,i}$  playing the role of  $\hat{\omega}_{i,i}$  — the test statistic  $t_T^{min}$  of (14) can be considered a feasible QLR test statistic under the assumption of a diagonal covariance matrix  $\boldsymbol{\Omega}$ . And hence the Cons test of Subsection 3.1 can be interpreted as a feasible QLR test for testing problem (9).

The assumption of  $\boldsymbol{\Omega}$  being diagonal does not generally hold true for our application. It is, therefore, natural to consider the following feasible QLR test statistic under an arbitrary covariance matrix  $\boldsymbol{\Omega}$ :

$$\inf_{\boldsymbol{\Delta} \in R_3^c} (\hat{\boldsymbol{\Delta}}_T - \boldsymbol{\Delta})' \hat{\boldsymbol{\Omega}}^{-1} (\hat{\boldsymbol{\Delta}}_T - \boldsymbol{\Delta}) , \quad (30)$$

where  $\hat{\boldsymbol{\Omega}}$  is a suitable estimator of  $\boldsymbol{\Omega}$ .

Unfortunately, the choice of test statistic (30) results in a test which can be liberal in finite samples; at least when  $\hat{\Omega}$  is the natural estimator of  $\Omega$ , namely the sample covariance matrix of the observations  $\mathbf{d}_1, \dots, \mathbf{d}_T$ .<sup>7</sup> This finding is in line with previous results in the literature that show that tests based on quadratic test statistics using the inverse of a sample covariance matrix are often liberal in finite samples; for example, see Hayashi (2000, Section 3.5), Ren and Shimotsu (2009), and the references therein.

## B Test based on minimum is UMP among monotone tests

As in Appendix A, assume  $\mathbf{X} \sim N(\boldsymbol{\mu}, \Omega)$ , where  $\Omega$  is a known positive definite matrix. Consider the testing problem (9) applied to the vector  $\boldsymbol{\mu}$ , namely

$$H_0 : \boldsymbol{\mu} \in R_1 \cup R_2 \quad \text{vs.} \quad H_1 : \boldsymbol{\mu} \in R_3 . \quad (31)$$

Restrict attention to monotone tests as described in Remark 3.1. Thus, if a test based on  $\mathbf{X}$  rejects  $H_0$ , then it must also reject  $H_0$  based on any  $\mathbf{X}' \geq \mathbf{X}$ .

**Theorem B.1** *Consider the above testing problem. Among monotone level  $\alpha$  (nonrandomized) tests, the test that rejects when  $\min_i X_i / \sqrt{\omega_{i,i}} \geq z_{1-\alpha}$  is uniformly most powerful (UMP), where  $z_{1-\alpha}$  is the  $1 - \alpha$  quantile of the standard normal distribution.*

PROOF: Let  $\mathcal{R} \subset \mathbb{R}^N$  be the rejection region of the claimed UMP test, so

$$\mathcal{R} \equiv \{\mathbf{x} : \min_i x_i / \sqrt{\omega_{i,i}} \geq z_{1-\alpha}\} . \quad (32)$$

Assume there exists another test with rejection region  $\mathcal{R}'$  that has better power against even one alternative  $\boldsymbol{\mu} \in R_3$ . Then, it cannot be the case that  $\mathcal{R}' \subseteq \mathcal{R}$ , or the test based on  $\mathcal{R}'$  could not have better power (as it would never reject unless the test based on  $\mathcal{R}$  did). Hence,  $\mathcal{R}'$  must include a point  $\mathbf{y} \equiv (y_1, \dots, y_N)' \notin \mathcal{R}$ . Therefore, for at least some  $j$ , it holds true that  $y_j < \sqrt{\omega_{j,j}} z_{1-\alpha}$ . Let

$$\mathcal{R}'' \equiv \{\mathbf{x} : \mathbf{x} \geq \mathbf{y}\} . \quad (33)$$

Since  $\mathcal{R}'$  is assumed monotone<sup>8</sup>, it must be the case that  $\mathcal{R}'' \subseteq \mathcal{R}'$ . But then, the supremum of the probability of a type 1 error of the test based on  $\mathcal{R}'$  satisfies

$$\sup_{\boldsymbol{\mu} \in R_1 \cup R_2} \mathbb{P}_{\boldsymbol{\mu}}\{\mathcal{R}'\} \geq \sup_{\boldsymbol{\mu} \in R_1 \cup R_2} \mathbb{P}_{\boldsymbol{\mu}}\{\mathcal{R}''\} . \quad (34)$$

<sup>7</sup>Results of corresponding Monte Carlo studies are not included in the paper, but are available from the authors upon request.

<sup>8</sup>A region  $\mathcal{R} \subseteq \mathbb{R}^N$  is called *monotone* (increasing) provided the following condition holds: if  $\mathbf{x} \in \mathcal{R}$  and  $\mathbf{x}' \geq \mathbf{x}$ , then necessarily  $\mathbf{x}' \in \mathcal{R}$  as well.

The right hand side is bounded below by  $\mathbb{P}_{\boldsymbol{\mu}}\{\mathcal{R}''\}$  with  $\boldsymbol{\mu}$  satisfying  $\mu_i = 0$  for some  $i$  and  $\mu_j = B$  for all  $j \neq i$ , where  $B$  is some large value. As  $B \rightarrow \infty$ , this probability becomes

$$\mathbb{P}_{\mu_i=0}\{X_i \geq y_i\} > \mathbb{P}_{\mu_i=0}\{X_i \geq \sqrt{\omega_{i,i}} \cdot z_{1-\alpha}\} = \alpha . \quad (35)$$

This is a contradiction, as the test would not be level  $\alpha$ .

The argument actually applies to any family of distributions  $F_{\boldsymbol{\theta}}$  that are stochastically increasing in a parameter  $\boldsymbol{\theta}$ , meaning  $\mathbb{P}_{\boldsymbol{\theta}_1}\{\mathcal{R}\} \leq \mathbb{P}_{\boldsymbol{\theta}_2}\{\mathcal{R}\}$  whenever  $\mathcal{R}$  is a monotone (increasing) region and  $\boldsymbol{\theta}_1 \leq \boldsymbol{\theta}_2$ . This holds true for any multivariate location model.

In particular, under our maintained assumptions, an approximate multivariate normal location model applies to the vector of test statistics  $\mathbf{t}_T = (t_{T,1}, \dots, t_{T,N})'$ , at least when the sample size  $T$  is large. And, therefore, in an approximate set-up, the Cons test is UMP among all monotone level  $\alpha$  tests for testing problem (9).

**Remark B.1** Consider the multivariate normal model of this section, specializing to the case where  $\boldsymbol{\Omega}$  is the identity matrix. Testing problem (7) applied to the vector  $\boldsymbol{\mu}$  becomes

$$H_0 : \boldsymbol{\mu} \in R_1 \quad \text{vs.} \quad H_1 : \boldsymbol{\mu} \in R_3 . \quad (36)$$

The exact finite sample analogue of the MR test of P&T in this case would be to reject when  $\min X_i \geq c_N(1 - \alpha)$ , where the critical value is determined by

$$\mathbb{P}_{\boldsymbol{\mu}=\mathbf{0}}\{\min_i X_i \geq c_N(1 - \alpha)\} = \alpha . \quad (37)$$

Therefore,  $[1 - \Phi(c_N(1 - \alpha))]^N = \alpha$ , or

$$c_N(1 - \alpha) = \Phi^{-1}(1 - \alpha^{1/N}) . \quad (38)$$

It follows that, if the correct null hypothesis is instead specified by (31), the probability of a type 1 error is maximized by maximizing

$$\mathbb{P}_{\boldsymbol{\mu}}\{\min X_i \geq c_N(1 - \alpha)\} \quad (39)$$

over  $\boldsymbol{\mu} \in R_1 \cup R_2$ . This probability is maximized when one  $\mu_i = 0$  and the remaining  $\mu_j$  tend to  $\infty$ . It follows that the maximum probability of a type 1 error of the MR test is exactly

$$\mathbb{P}_{\mu_1=0}\left\{X_1 \geq \Phi^{-1}(1 - \alpha^{1/N})\right\} = 1 - \Phi\left[\Phi^{-1}(1 - \alpha^{1/N})\right] = \alpha^{1/N} , \quad (40)$$

which is bigger than  $\alpha$ , unless  $N = 1$ , and it approaches one as  $N$  increases. Clearly, such a test does not control the type 1 error for our formulation (31) of the testing problem and is much too liberal. ■



## C Joint confidence region for $\Delta$

This section briefly summarizes the single-step method described in [Romano and Wolf \(2005\)](#) to compute the upper limits  $\hat{\mathbf{u}}_T \equiv (\hat{u}_{T,1}, \hat{u}_{T,2}, \dots, \hat{u}_{T,N})'$  utilized in Subsection 3.3. Specifically, they are of the form

$$\hat{u}_{T,i} \equiv \hat{\Delta}_{T,i} + \hat{d}_1 \cdot \hat{\sigma}_{T,i} . \quad (41)$$

Here,  $\hat{d}_1$  is a (consistent) estimate of  $d_1$ , which is defined as the  $1 - \beta$  quantile of the sampling distribution of the random variable

$$\max_i \frac{(\hat{\Delta}_{T,i} - \Delta_i)}{\hat{\sigma}_{T,i}} . \quad (42)$$

The following algorithm details how  $d_1$  is approximated via the bootstrap.

### Algorithm C.1 (Computation of $\hat{d}_1$ )

- Generate bootstrap data  $\{\mathbf{r}_1^*, \mathbf{r}_2^*, \dots, \mathbf{r}_T^*\}$  and compute statistics  $\hat{\Delta}_{T,i}^*$  and  $\hat{\sigma}_{T,i}^*$  from these data, for  $i = 1, \dots, N$
- Compute  $\max_T^* \equiv \max_i (\hat{\Delta}_{T,i}^* - \hat{\Delta}_{T,i}) / \hat{\sigma}_{T,i}^*$
- Repeat this process  $B$  times, resulting in statistics  $\max_{T,1}^*, \dots, \max_{T,B}^*$
- The  $1 - \beta$  empirical quantile of these  $B$  statistics is the bootstrap estimate  $\hat{d}_1$

For motivation and further details, the reader is referred to [Romano and Wolf \(2005, Section 4\)](#).

## D Proofs

**PROOF OF THEOREM 4.1:** Denote the correlation matrix corresponding to the limiting covariance matrix  $\mathbf{\Omega}$  of Assumption (A1) by  $\mathbf{\Pi}$ , with typical element  $\pi_{i,j}$ , that is

$$\pi_{i,j} = \frac{\omega_{i,j}}{\sqrt{\omega_{i,i} \cdot \omega_{j,j}}} . \quad (43)$$

Define centered statistics by

$$v_{T,i} \equiv \frac{\hat{\Delta}_{T,i} - \Delta_i}{\hat{\sigma}_{T,i}} , \quad (44)$$

and let  $\mathbf{v}_T \equiv (v_{T,1}, \dots, v_{T,N})'$ . Then  $\mathbf{v}_T$  converges in distribution to  $N(\mathbf{0}, \mathbf{\Pi})$ . Let  $\mathbf{y} \equiv (y_1, \dots, y_N)'$  be a random variable with distribution  $N(\mathbf{0}, \mathbf{\Pi})$  and let  $y^{\min} \equiv \min_i y_i$ . Also, denote by  $c_{1-\alpha}$  the  $1 - \alpha$  quantile of the distribution of  $y^{\min}$ .

Furthermore, since

$$t_{T,i} \equiv \frac{\hat{\Delta}_{T,i}}{\hat{\sigma}_{T,i}} = v_{T,i} + \frac{\Delta_{T,i}}{\hat{\sigma}_{T,i}} , \quad (45)$$

it holds that

$$t_{T,i} \text{ converges in distribution to } \begin{cases} -\infty & \text{if } \Delta_i < 0 \\ N(0, 1) & \text{if } \Delta_i = 0 \\ \infty & \text{if } \Delta_i > 0 \end{cases} . \quad (46)$$

(Here, in the first case and in the third case, convergence in distribution is actually equivalent to convergence in probability.)

To prove part (ia), first consider the case  $\Delta = \mathbf{0}$ . It then follows that  $t_T^{min}$  converges in distribution to  $y^{min}$  and that the critical value of the test converges to  $c_{1-\alpha}$  in probability. Hence, the rejection probability converges to  $\alpha$ . Any other parameter  $\Delta$  considered satisfies  $\min_i \Delta_i < 0$ , in which case  $t_T^{min}$  converges to  $-\infty$  in probability, resulting in a limiting rejection probability of zero.

To prove part (ib), any parameter  $\Delta$  considered satisfies  $\min_i \Delta_i = 0$  and  $\max_i \Delta_i > 0$ . Let  $I(\Delta) \equiv \{i : \Delta_i = 0\}$ . Then  $t_T^{min}$  converges in distribution to  $y_{I(\Delta)}^{min} \equiv \min_{i \in I(\Delta)} y_i$ . Note that  $y_{I(\Delta)}^{min}$  is stochastically larger than  $y^{min}$ , since  $I(\Delta)$  is a strict subset of  $\{1, \dots, N\}$ . The rejection probability thus converges to  $\mathbb{P}\{y_{I(\Delta)}^{min} > c_{1-\alpha}\} > \mathbb{P}\{y^{min} > c_{1-\alpha}\} = \alpha$ .

To prove part (ic), any parameter  $\Delta$  considered satisfies  $\min_i \Delta_i > 0$ , in which case  $t_T^{min}$  converges to  $\infty$  in probability, resulting in a limiting rejection probability of one.

To prove part (iia), any parameter  $\Delta$  considered satisfies  $\min_i \Delta_i \leq 0$ . To start out, consider the Cons test. The critical value converges to  $z_{1-\alpha}$  in probability. But, in the limit, the distribution of the test statistic  $t_T^{min}$  is (weakly) stochastically smaller than the standard normal distribution for all values of  $\Delta$  considered. So the limiting rejection probability of the test cannot exceed  $1 - \alpha$ . Next, consider the CE test. Since  $\hat{\Delta}_T$  converges in to  $\Delta$  in probability, the following is not difficult to show. First, if  $\min_i \Delta_i < 0$ , then the limiting rejection probability is zero; see Remark 3.1. Second, if  $\Delta$  contains one zero entry and the remaining entries are positive, the critical value converges to  $z_{1-\alpha}$  in probability and the limiting rejection probability is equal to  $\alpha$ . Third, if  $\Delta$  contains several zero entries and the remaining (if any) entries are positive, let again  $I(\Delta) \equiv \{i : \Delta_i = 0\}$ . The test statistic  $t_T^{min}$  converges in distribution to  $\min_{i \in I(P)} y_i$ . Denote the  $1 - \alpha$  quantile of the distribution of  $\min_{i \in I(P)} y_i$  by  $c_{I(P), 1-\alpha}$ . Then, for any  $\epsilon > 0$ , the probability that the critical value of the test will be smaller than  $c_{I(P), 1-\alpha} - \epsilon$  converges to zero. As a result, the limiting rejection probability of the test cannot exceed  $\alpha$ . Finally, consider the Two-Step test. Denote by  $E$  the set in the sample space on which the joint confidence region (18) contains the true parameter  $\Delta$ . Hence, on the set  $E$ , the employed critical value is (weakly) larger compared to computing the critical value under the true parameter. As a result, the limiting rejection probability cannot exceed the nominal level of the test, which is  $\alpha - \beta$ . Assume without loss of generality that on the set  $E^c$  (that is, on the complement of the set  $E$ ), the test always rejects. By design of the joint confidence region (18), the limiting probability of  $E^c$  cannot exceed  $\beta$ . In sum, the

limiting rejection probability cannot exceed  $\alpha - \beta + \beta = \alpha$ .

To prove part (iib), any parameter  $\Delta$  considered satisfies  $\min_i \Delta_i > 0$ , in which case  $t_T^{\min}$  converges to  $\infty$  in probability, resulting in a limiting rejection probability of one. Note here that for all three tests considered, the critical value converges to  $z_{1-\alpha}$  in probability. ■

## References

- Andrews, D. W. K. (1991). Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica*, 59:817–858.
- Andrews, D. W. K. and Monahan, J. C. (1992). An improved heteroskedasticity and autocorrelation consistent covariance matrix estimator. *Econometrica*, 60:953–966.
- Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, 7:1–26.
- Fama, E. F. (1984). Term premium in bond returns. *Journal of Financial Economics*, 13:529–546.
- Hall, P. (1992). *The Bootstrap and Edgeworth Expansion*. Springer, New York.
- Hansen, P. R. (2003). Asymptotic tests of composite hypotheses. Economics Working Paper No. 03-09, Brown University. Available at SSRN: <http://ssrn.com/abstract=399761>.
- Hansen, P. R. (2005). A test for superior predictive ability. *Journal of Business and Economics Statistics*, 23:365–380.
- Hayashi, F. (2000). *Econometrics*. Princeton University Press, Princeton.
- Lahiri, S. N. (2003). *Resampling Methods for Dependent Data*. Springer, New York.
- Loh, W. (1985). A new method for testing separate families of hypotheses. *Journal of the American Statistical Association*, 80:362–368.
- Owen, A. B. (2001). *Empirical Likelihood*. Chapman & Hall/CRC, Boca Raton.
- Patton, A. J. and Timmermann, A. (2010). Monotonicity in asset returns: New tests with applications to the term structure, the CAPM, and portfolio sorts. *Journal of Financial Economics*, 98:605–625.
- Perlman, M. D. and Wu, L. (1999). The emperor’s new tests (with discussion). *Statistical Science*, 14(4):355–381.
- Politis, D. N. and Romano, J. P. (1992). A circular block-resampling procedure for stationary data. In LePage, R. and Billard, L., editors, *Exploring the Limits of Bootstrap*, pages 263–270. John Wiley, New York.
- Politis, D. N. and Romano, J. P. (1994). The stationary bootstrap. *Journal of the American Statistical Association*, 89:1303–1313.

- Ren, Y. and Shimotsu, K. (2009). Improvement in finite sample properties of the Hansen-Jagannathan distance test. *Journal of Empirical Finance*, 16:483–506.
- Romano, J. P. and Wolf, M. (2000). Finite sample nonparametric inference and large sample efficiency. *Annals of Statistics*, 28(3):756–778.
- Romano, J. P. and Wolf, M. (2005). Stepwise multiple testing as formalized data snooping. *Econometrica*, 73(4):1237–1282.
- Scherer, B. and Martin, R. D. (2005). *Introduction to Modern Portfolio Optimization*. Springer, New York.
- Wolak, F. A. (1987). An exact test for multiple inequality and equality constraints in the linear regression model. *Journal of the American Statistical Association*, 82:782–793.
- Wolak, F. A. (1989). Testing inequality constraints in linear econometric models. *Journal of Econometrics*, 31:205–235.

Small Size, $t_T^{min} = 2.02$		
Test	MR	Cons
Critical Value	0.20	1.60

Mid Size, $t_T^{min} = 1.33$		
Test	MR	Cons
Critical Value	0.13	1.67

Large Size, $t_T^{min} = 0.10$		
Test	MR	Cons
Critical Value	0.17	1.88

Table 1: Test results for equity returns corresponding to the three panels of Figure 8. The nominal level is  $\alpha = 0.05$ . All return differentials were multiplied by negative one before feeding the data to the two testing procedures.

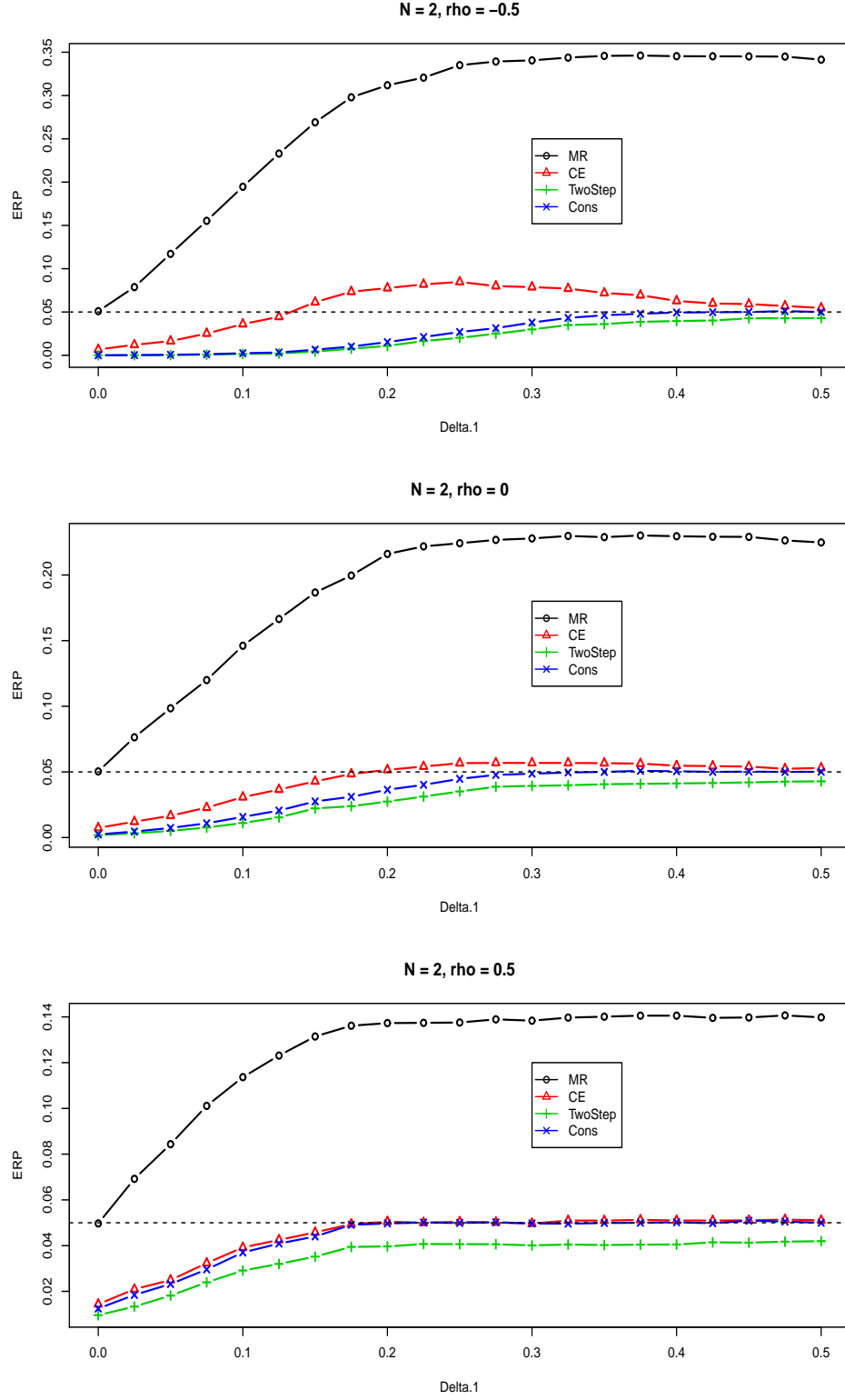


Figure 1: Empirical rejection probabilities (ERPs) for various tests in dimension  $N = 2$  as a function of  $\Delta_1$ , where  $\Delta = (\Delta_1, 0)$ .

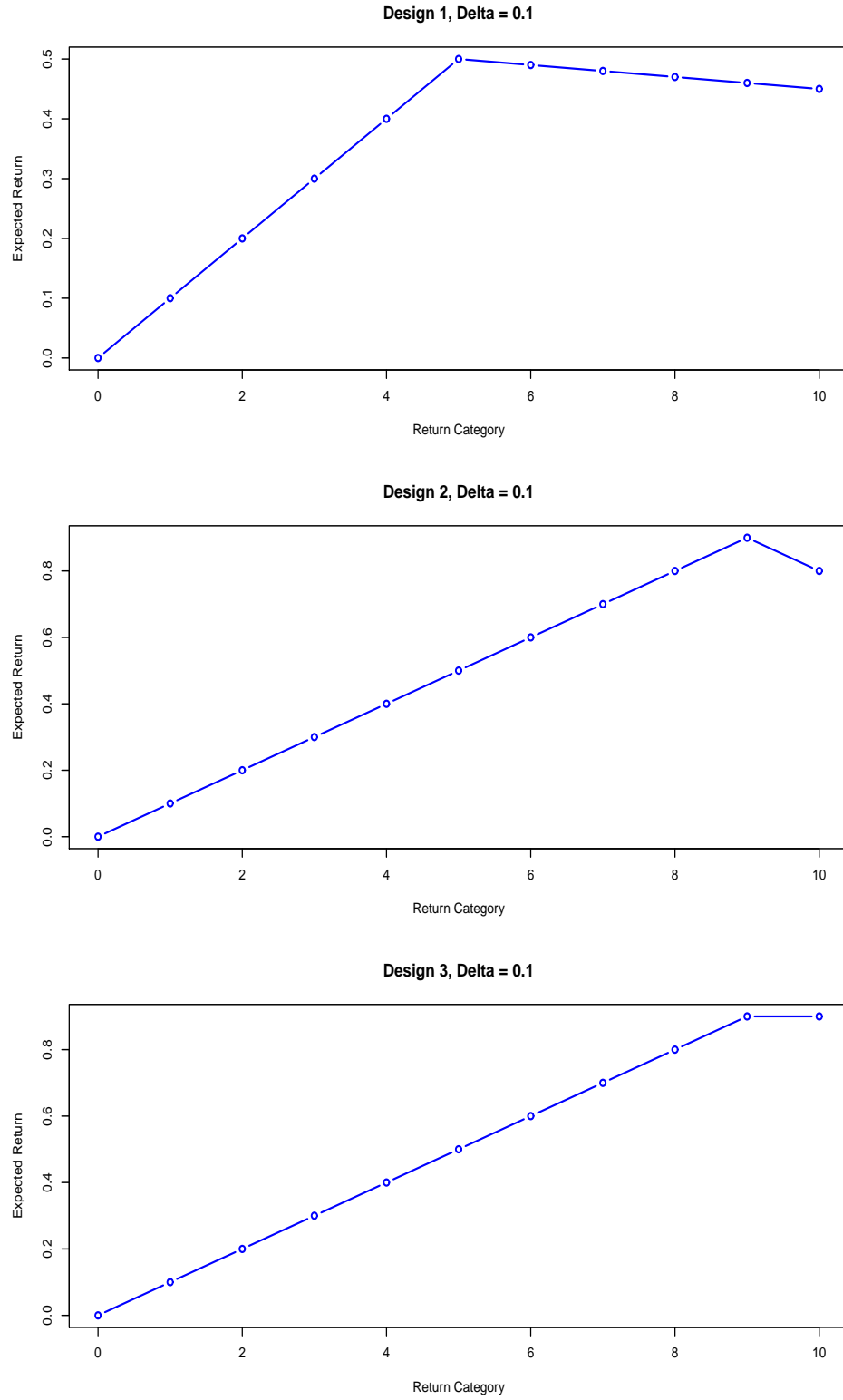


Figure 2: Vectors of expected returns, for  $\Delta = 0.1$ , corresponding to the three null designs detailed in Subsection 5.2.



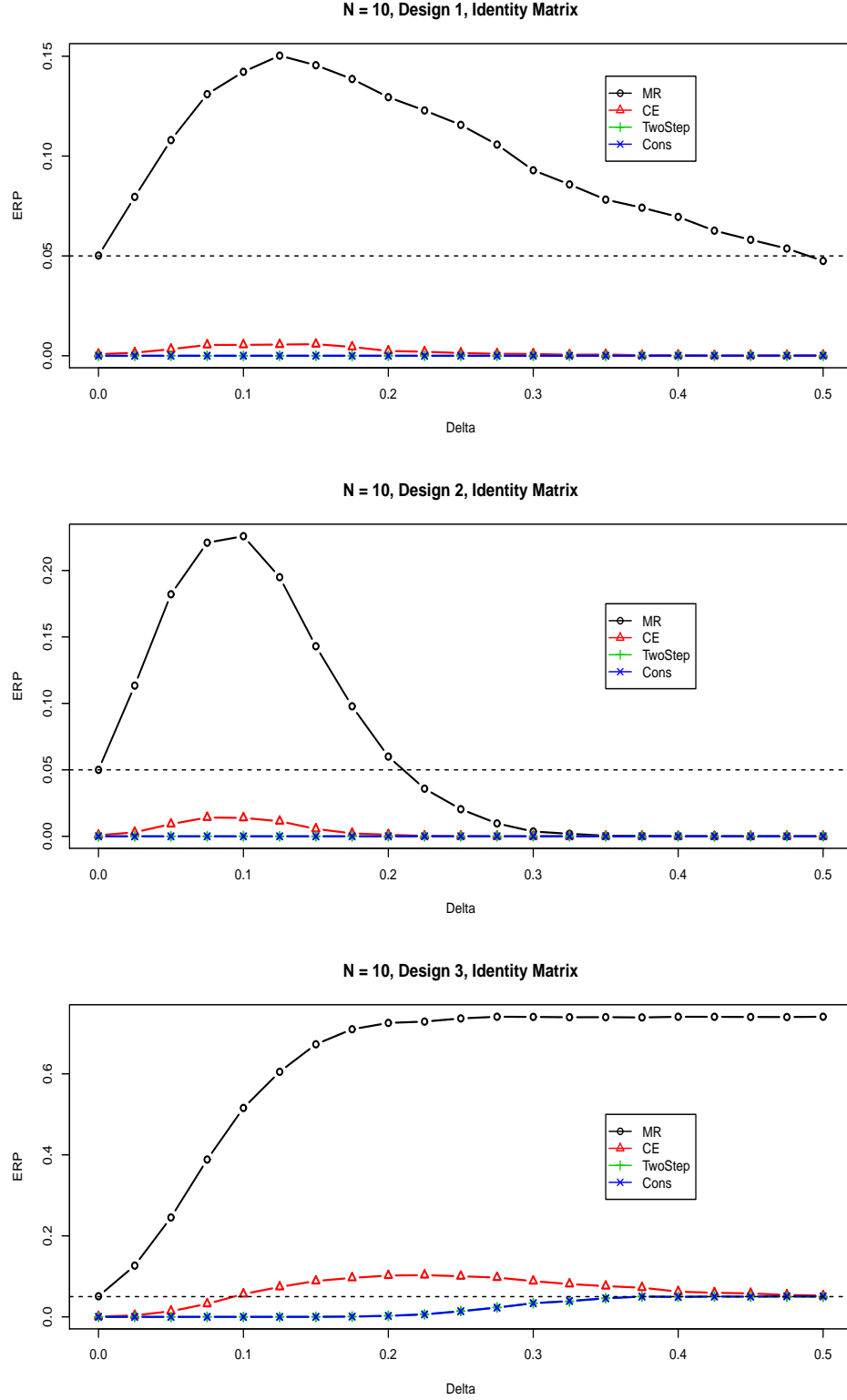


Figure 3: Empirical rejection probabilities (ERPs) under  $H_0$  for various tests in dimension  $N = 10$  as a function of  $\Delta$ , where the three designs are detailed in Subsection 5.2. The covariance matrix of the return differentials is the identity matrix.

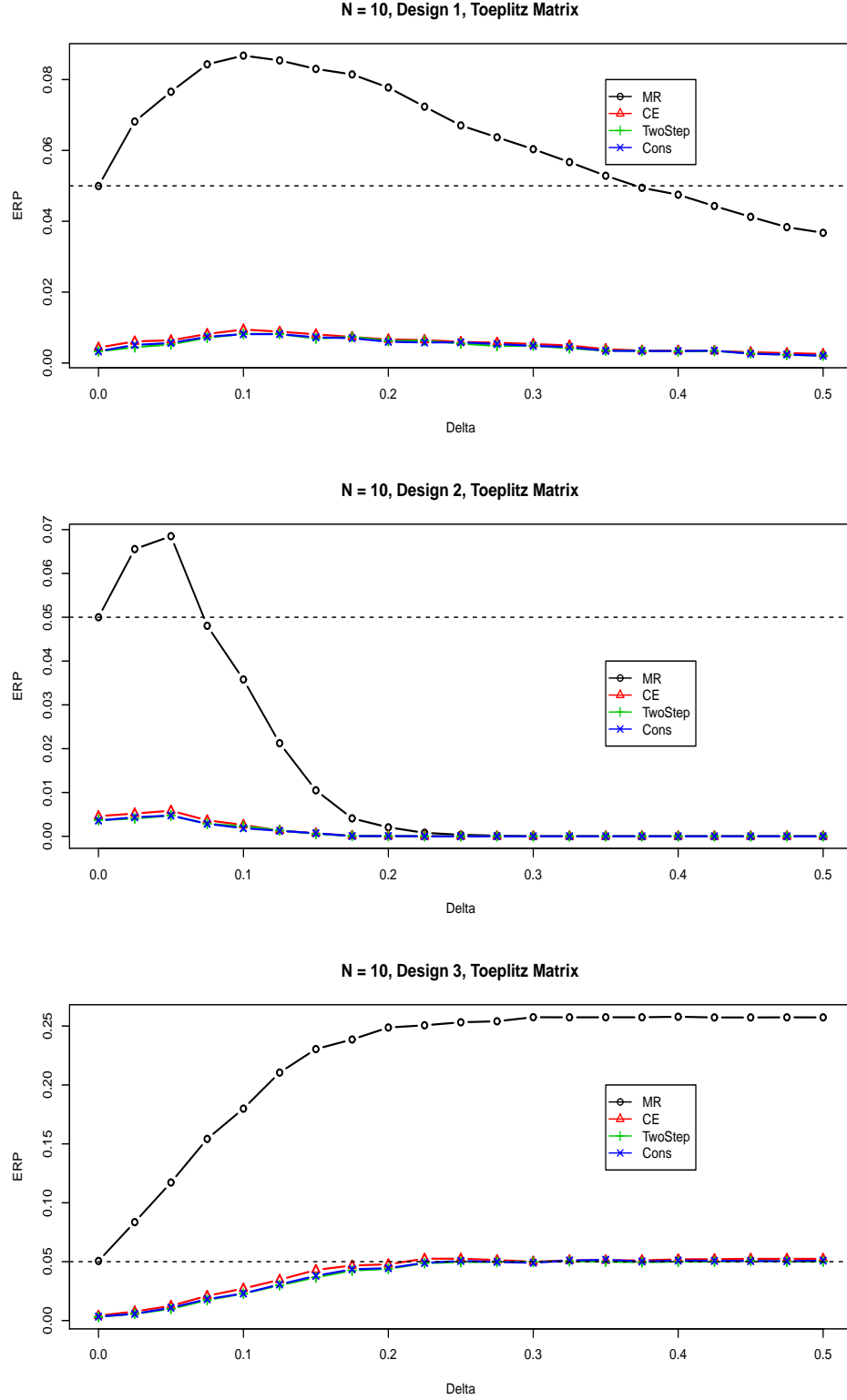


Figure 4: Empirical rejection probabilities (ERPs) under  $H_0$  for various tests in dimension  $N = 10$  as a function of  $\Delta$ , where the three designs are detailed in Subsection 5.2. The covariance matrix of the return differentials is a Toeplitz matrix.

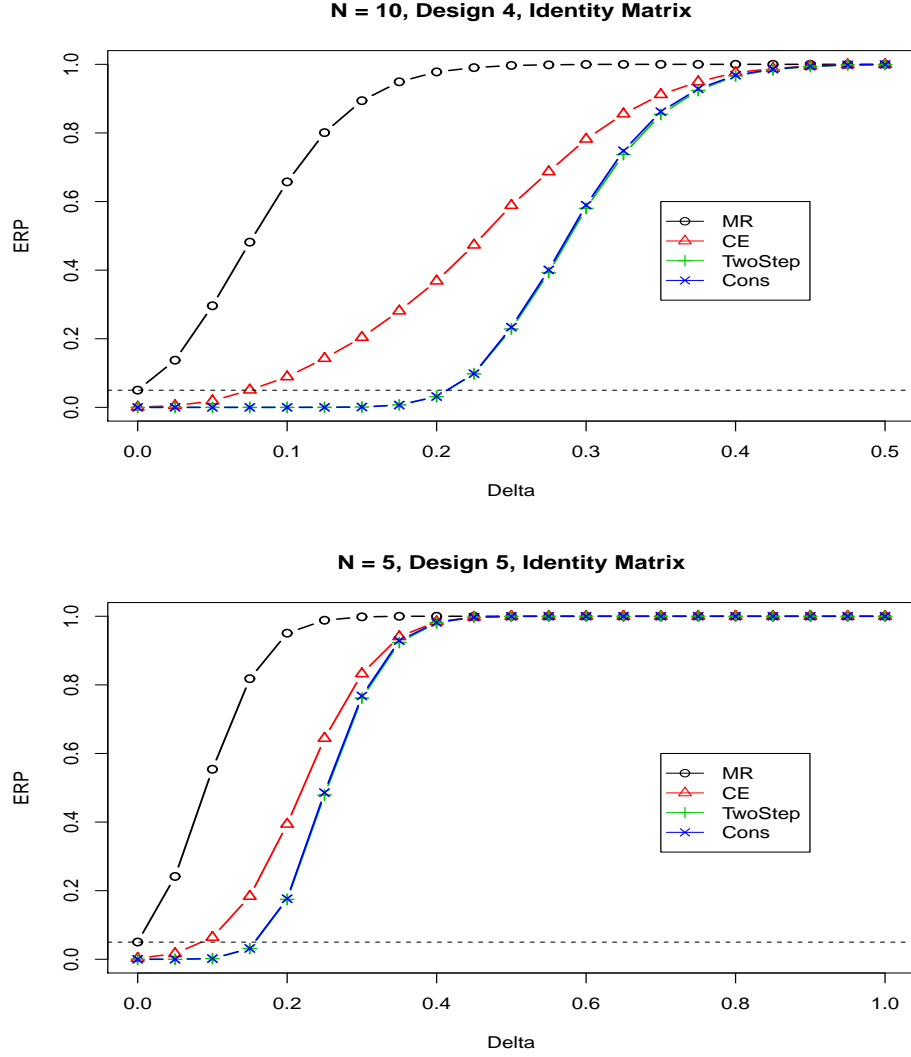


Figure 5: Empirical rejection probabilities (ERPs) under  $H_1$  for various tests in dimensions  $N = 10$  and  $N = 5$  as a function of  $\Delta$ , where the two designs are detailed in Subsection 5.3. The covariance matrix of the return differentials is the identity matrix.

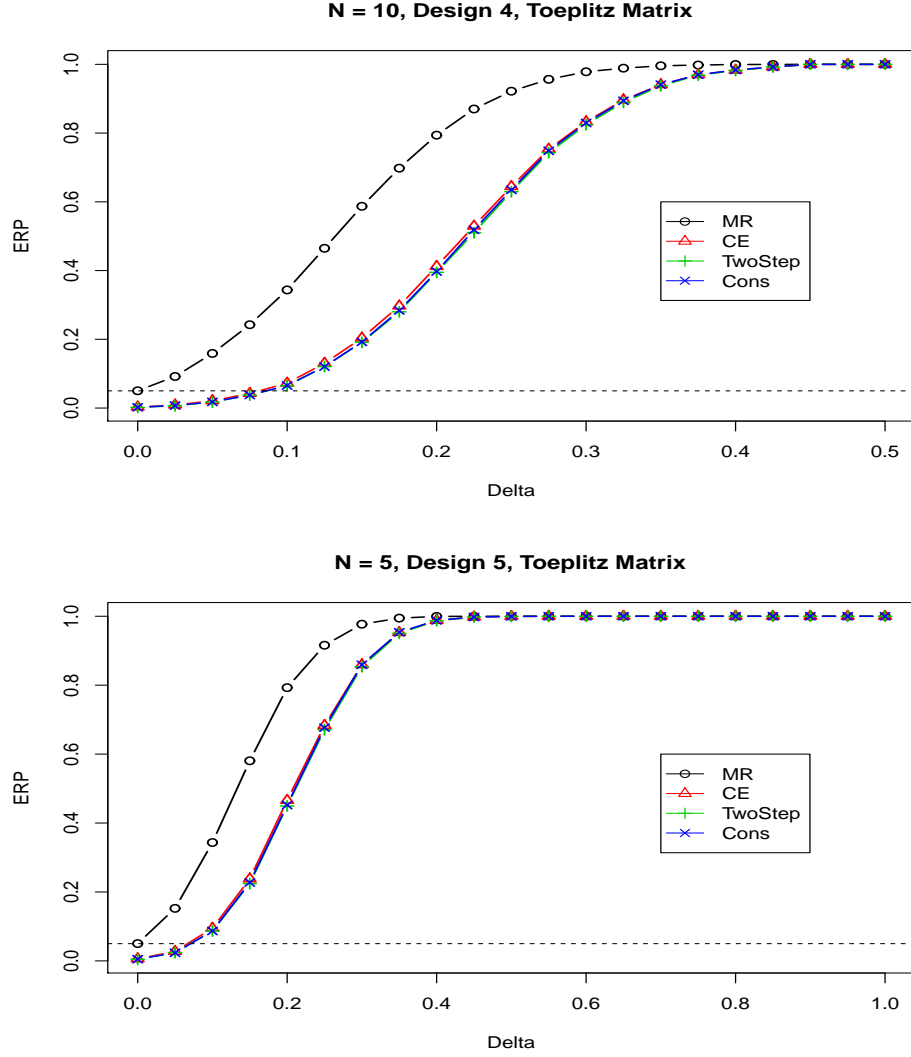


Figure 6: Empirical rejection probabilities (ERPs) under  $H_1$  for various tests in dimensions  $N = 10$  and  $N = 5$  as a function of  $\Delta$ , where the two designs are detailed in Subsection 5.3. The covariance matrix of the return differentials is a Toeplitz matrix.

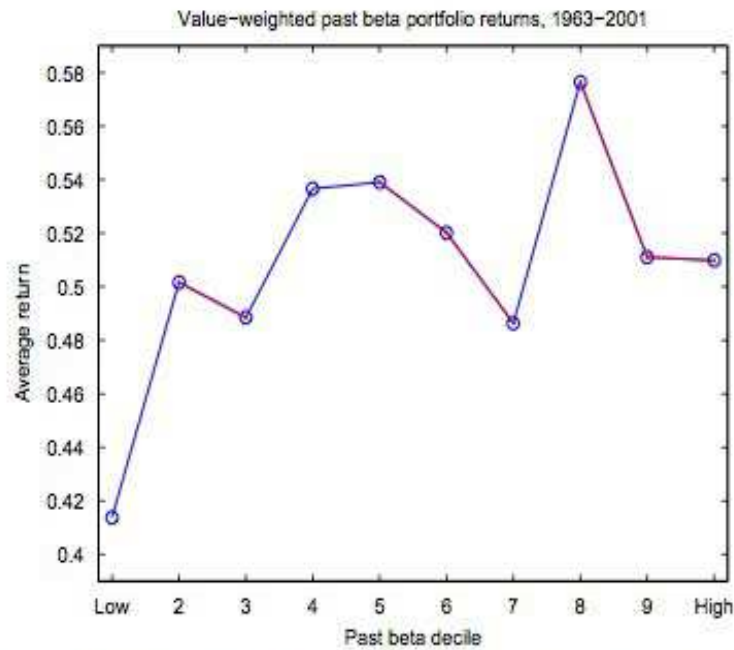


Fig. 1. Average monthly returns on decile portfolios formed on past 12-month Capital Asset Pricing Model (CAPM) beta, from July 1963 to December 2001.

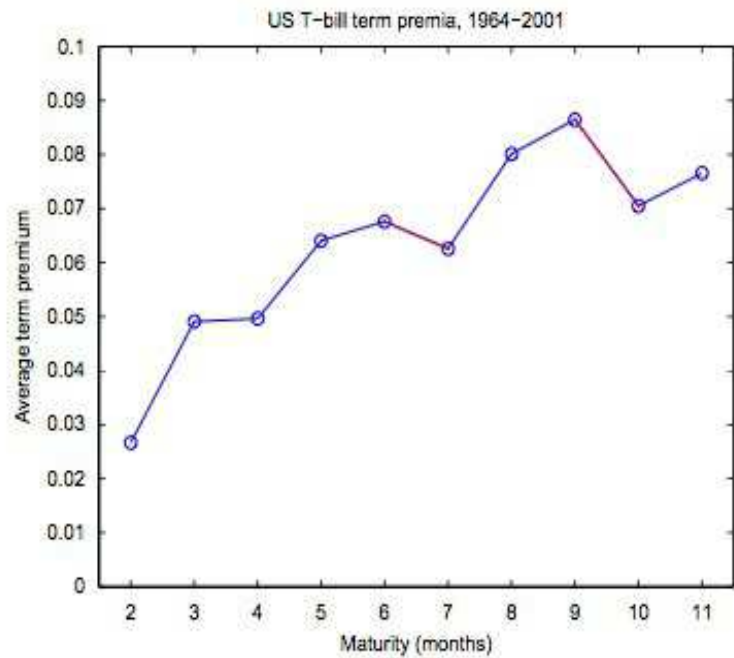


Fig. 2. Average monthly term premia for US T-bills, relative to a T-bill with one month to maturity, over the period January 1964 to December 2001.

Figure 7: Reproduction of Fig. 1 and Fig. 2 of [Patton and Timmermann \(2010\)](#).

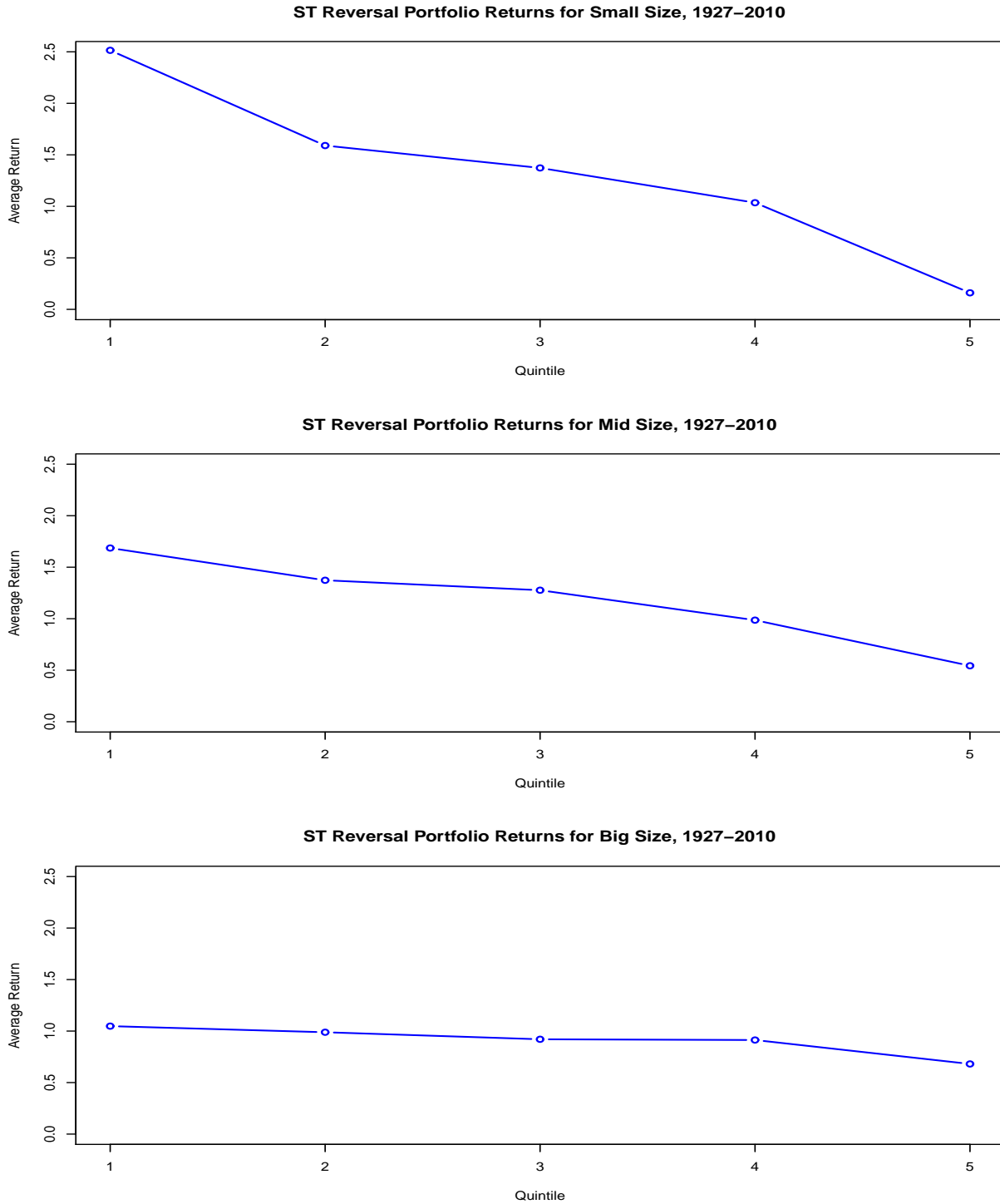


Figure 8: The upper panel displays average monthly value-weighted returns on quintile portfolios formed on short-term reversal for small-size firms, from January 1927 to December 2010. Similar for the middle and lower panel; except that mid-size firms and big-size firms, respectively, are used instead.